2.3 PUBLIC MULTILINGUAL KNOWLEDGE MANAGEMENT INFRASTRUCTURE FOR THE DIGITAL SINGLE MARKET (2016.16)

2.3.1 IDENTIFICATION OF THE ACTION

Type of Activity	Creation of a public multilingual knowledge infrastructure
Service in charge	Publications Office of the European Union
Associated Services	DG Connect DG DIGIT DG DGT European Parliament, DG TRAD, Terminology Coordination Centre de Traduction

2.3.2 EXECUTIVE SUMMARY

In their open letter to the European Commission the European Language (Technology) Community claims: Europe’s Digital Single Market must be multilingual![16] EU cross-border online services represent only 4% of the global Digital Market and only 7% of small and medium sized enterprises (SMEs) in the EU are actually selling cross-border.[17] Providing support for the EU economy and in particular to SMEs to overcome the language barriers will help to unlock the e-Commerce potential within the EU.

The objective of this action is to support enterprises and in particular the language technology industry with the implementation of the necessary multilingual tools and features in order to improve cross border accessibility of e-Commerce solutions. It will also be an input to the CEF Automated Translation Platform, a common building block implemented through the CEF programme to be used by European cross-border online public services.

The public multilingual knowledge infrastructure will be governed by a representative subgroup of stakeholders of the final system.

In this context multilingual tools and features refer to capabilities as machine translation, localisation and multilingual search. The public multilingual knowledge infrastructure should reduce the investments of enterprises for the creation of their individual knowledge management systems by providing an agreed, open, reliable and persistent public core knowledge management system. This would also create space for innovation instead of wasting resources for redundant activities.

Public administrations and public entities within the EU will largely benefit from this initiative, in particular regarding the internationalisation of their e-Services. They will be able to share and to

[16] See https://ec.europa.eu/futurium/en/content/european-language-and-language-technology-community-europes-digital-single-market-must-be

[17] See http://europa.eu/rapid/attachment/IP-15-4653/en/Digital_Single_Market_Factsheet_20150325.pdf

valorise existing taxonomies/terminologies and to extend their multilingual capabilities. This will also help to increase the interoperability between public administrations within EU in general.

To realise the public multilingual knowledge infrastructure the following aspects need to be addressed:

- Implementation of a technical infrastructure to expose existing multilingual taxonomies/terminologies in a standardised way based on semantic technology and Semantic Web standards;
- Implementation of existing alignments between terminologies and creation of further alignments and relations in order to enable interoperability;
- Creation and maintenance of meaningful supplements, i.e. of terms and relations that complete the coverage of the multilingual knowledge infrastructure and facilitate interoperability;
- Set-up of a community and a governance structure to extend systematically the coverage of the core infrastructure by the integration of supplementary public multilingual taxonomies/terminologies.

The cornerstone of the public multilingual knowledge infrastructure will be EuroVoc, the multilingual, multidisciplinary thesaurus covering the activities of the EU, which is managed by the Publications Office. The project will also benefit from already existing alignments of EuroVoc with other thesauri (Agrovoc, Eclas, Gemet, Anubis and Inspire[18]).

In addition, it should be investigated how the publication of the information as Linked Open Data (LOD) could be enhanced by introducing lexical semantic links between semantically equivalent and similar entities in an automatized way.

2.3.3 OBJECTIVES

The objectives of the proposed activity are:

- To provide an agreed, reliable, persistent and extensible public multilingual terminology platform for multiple purposes and for multiple stakeholders composed of open public multilingual resources.
- To increase the interoperability of existing multilingual terminologies, in particular by aligning and linking them with other existing terminologies. Linking will enable at the same time specialisation (for example by linking a concept of a more general taxonomy/terminology with the corresponding concept of a domain specific taxonomy/terminology) and broadening (for example by linking similar concepts at the same level of granularity).
- To establish the initial governance structure to support and to supervise the execution of the project as well as the implementation, the management and the evolution of the final

[18] See https://open-data.europa.eu/en/data/dataset/eurovoc

system. Synergy with the governance structure, which is being set-up for the CEF.AT platform (see http://www.lr-coordination.eu/anchor-points) will be established.

- To contribute to the further standardisation of data models for thesauri and lexical knowledge database representations using latest semantic technologies.
- To support the contributing institutions with the transformation of their resources into adopted semantic format of the platform.
- To further develop the LOD capabilities of the system, in particular by enhancing the creation of semantic links between similar and related concepts.

2.3.4 SCOPE

The objective of the proposed activity in the scope of the ISA² programme is to verify the feasibility of the approach and to prepare the technical and the organisation aspects for the definitive and permanent implementation of an open public multilingual knowledge infrastructure managed by the EU Institutions.

Nevertheless, it will deliver already results, which could be used and applied by public administrations and bodies of Member States and EU Institutions independent from the public multilingual knowledge infrastructure project itself.

In scope

1. Adoption of a standard representation for multilingual terminologies (candidates include SKOS[19], Lemon[20]…)
2. Definition of a core data model based on the standard representation 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, which will be made available through the platform
3. Specification of the technical architecture of the public multilingual knowledge infrastructure and the necessary services to access and to manage the system
4. Proof of concept – implementation of a first operational release of the system to demonstrate the core services of the system
5. Set-up/adoption of an adequate initial governance structure
6. Definition of an iterative implementation strategy, i.e. the specifications and the roadmap for the extension of the initial release of the system into a public service, which will be managed, further developed and maintain by the EU Institutions and governed by all contributors, i.e. in particular public entities within in EU. The result of the proof of concept should be reused for the implementation of the final system.
7. Feasibility study in order to analyse and to test the creation of lexical semantic links between semantically equivalent and similar entities in an automatized way.

Out of scope

[19] See http://www.w3.org/2004/02/skos/

[20] See http://lemon-model.net/index.php

1. Implementation of the definitive, permanent platform, made available as a public service and free of charge.
2. Management, further development and maintenance of the definitive system.

2.3.5 PROBLEM STATEMENT

"Linguistic diversity is and must remain a cornerstone and treasured cultural asset of Europe. However, the language barriers created by our 24 official EU languages cause the European market to be fragmented and to fall short of its economic potential. Almost half of European citizens never shop online in languages other than their native tongue, access to public e-services is usually restricted to national languages and the richness of EU educational and cultural content is confined within linguistic communities. European SME's are at particular disadvantage, because the cost of providing services in multiple languages is prohibitive and has a negative impact on their competitiveness."[21]

This challenge needs to be addressed and a public multilingual knowledge infrastructure will contribute to reduce and to secure the investments of the different stakeholders in cross-border e-commerce solutions and related services because part of their implementation could really on an agreed, authentic and persistent set of multilingual terminology.

Because the contributions for public multilingual knowledge infrastructure will come from different stakeholders (essentially public administrations and bodies of EU Member States, EU Institutions and international organisations) the challenge is to build a system, which empowers the stakeholders to manage the development and evolution of their taxonomies/terminologies on an individual base, but at the same time enables interoperability through alignment and linking.

The only constraint should be that stakeholders have either to adopt the core data model proposed by the public multilingual knowledge infrastructure or, at least, have to be able to perform the necessary transformations to provide new releases in compliance with the core data model. Ideally, new releases have to be provided in a way that they can be integrated largely automatically.

[21] See https://ec.europa.eu/futurium/en/content/european-language-and-language-technology-community-europes-digital-single-market-must-be

2.3.6 EXPECTED BENEFICIARIES AND ANTICIPATED BENEFITS

Beneficiaries	Anticipated benefits
EU economy	Many studies have already between conducted to evaluate the possible economic impact of an increase in cross-border e-commerce between EU Member States. The creation of a real EU digital single market has become a priority of the Commission. The initiative will provide a contribution on the technological level. It will help to reduce the localisation effort for e-commerce platforms, enhance the quality of the domain specific terminology and improve their harmonisation. It will also facilitate the implementation of multilingual search capabilities.
EU language technology industry	Cost reduction and faster implementation of services related to cross-border e-commerce (machine-translation, localisation software, cross-language search solutions…). This will also increase the usability and searchability of resources for the creation of new, innovate services.
EU Member States	Will benefit in the context of the internationalisation of their e-government services for example to facilitate foreign investments in the local market. Will be able to improve interoperability with other Member States and/or public entities based on shared or aligned taxonomies/terminologies.
EU Institutions	Valorisation of existing multilingual taxonomies/terminologies, spin-offs for EU translation services and other multilingual services. It will help to increase the interoperability of multilingual LOD, which are made available by the EU Institutions.

2.3.7 RELATED EU ACTIONS / POLICIES

Action / Policy	Description of relation, inputs / outputs
Digital Single Market	The slogan of the initiative is "Bringing down barriers to unlock online opportunities". The action contributes to overcome the language barriers within the EU.
ISA action 1.1	ISA action 1.1 is about promoting semantic interoperability amongst the EU Member States. The proposed initiative proposes to deepen the semantic interoperability in the domain of terminology and increase its impact by including non-governmental stakeholders (language technology industry, e-commerce companies…).
CEF.AT platform	CEF.AT platform is aimed to deploy mature language technologies for the public sector and public online services (see: https://joinup.ec.europa.eu/community/cef/og_page/catalogue-building-blocks#AT).

2.3.8 REUSE OF SOLUTIONS DEVELOPED BY ISA, ISA^2 OR OTHER EU / NATIONAL INITIATIVES

The action could really on existing work of the Publications Office in the domain of multilingual thesauri, multilingual control vocabularies and Linked Open Data (LOD), in particular in the scope of EuroVoc.

The EuroVoc dataset and the controlled vocabularies managed by the Publications Office are already today available in semantic format (SKOS)[22] on the EU Open Data Portal and on the Publications Office's Metadata Registry[23]. The data is also partially exposed as LOD on the Web (through the Publications Office's common repository, the Cellar).

The results of the current ISA action 1.1 "Improving semantic interoperability in European eGovernment systems" should be taken into account and the results should be reused wherever possible.

As described in chapter 1.1.7, a close collaboration with existing eTranslation initiatives, in particular regarding the CEF Automated Translation (AT) building blocks, has to be established.

2.3.9 EXPECTED RE-USABLE OUTPUTS (solutions and instruments)

Output name	Core data model for multilingual taxonomies/terminologies
Description	Formal definition of the core data model for multilingual taxonomies/terminologies and its necessary extensions that will be implemented by the public multilingual knowledge infrastructure. The approach should be flexible in the way that data providers would be able to define private extensions, which would allow the upload of supplementary data that is available on their side and that could be useful for re-users. The aspects "provenance" and "license" have also to be taken into account.
Reference	PUB_MUL_TERM_FORMAT
Target release date / Status	Q1/2017

Output name	Technical architecture

[22] See http://open-data.europa.eu/en/data/dataset/eurovoc

[23] See http://publications.europa.eu/mdr/

Description	Technical design of the public multilingual knowledge infrastructure architecture including definition of all related services (ingestion of and access to data (including search), management of the infrastructure itself...).
Reference	PUB_MUL_TERM_ARCHITECTURE
Target release date / Status	Q2/2017

Output name	Governance structure
Description	Proposal for an adequate governance structure for the supervision of the public multilingual knowledge infrastructure.
Reference	PUB_MUL_TERM_ARCHITECTURE
Target release date / Status	Q2/2017

Output name	First release of the system
Description	Implementation of a first release of the system (repository and core services), which should be considered first of all as an operational proof of concept to demonstrate the core services of the platform and which will be reused to build the final system.
Reference	PUB_MUL_TERM_POC
Target release date / Status	Q2/2018

Output name	Implementation strategy
Description	Proposal of an iterative implementation strategy in order to prepare the political decision about whether the EU institutions will support the implementation of the public multilingual knowledge infrastructure and, if yes, how the system should be managed and financed.
Reference	PUB_MUL_TERM_STRATEGY
Target release date / Status	Q3/2018

Output name	Community building
Description	Proposal for implementation and organisation of a community (contributors, users…) to drive the further evolution of the system.
Reference	PUB_MUL_TERM_COMMUNITY
Target release date / Status	Q4/2018

Output name	Semantic links
Description	Feasibility study and prototype in order to explore the possibilities to enhance the semantic capabilities of the platform, in particular regarding the creation of lexical semantic links between semantically equivalent and similar entities in an automated way
Reference	PUB_MUL_TERM_SEMANTIC
Target release date / Status	Q2/2017

2.3.10 ORGANISATIONAL APPROACH

2.3.10.1 Expected stakeholders and their representatives

Stakeholders	Representatives
EU Institutions	European Parliament DG TRAD, Terminology Coordination Commission DIGIT DG DGT DG CONNECT Publications Office of the EU Centre de Traduction
International organisations	FAO…
Member states	
Language technology industry	
Research community	

2.3.10.2 Communication plan

The following table presents a first rough idea of a communication plan based of the different beneficiaries/interest groups, which have been identified in a first phase.
The existing platforms of the ISA programme in the domain of language technology will be taken into account for the set-up of an adequate communication platform.

Beneficiaries	Communication channel	Frequency
EU economy	Web (information about the activity on the ISA² website, publicity on the Publications Office and other EU Institutions websites)	Regular updates during the lifetime of the project.
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)	Regular updates during the lifetime of the project. 1 to 5 conferences per year
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)	Regular updates during the lifetime of the project. 1 to 3 workshops per year
EU Institutions	Meetings Workshop (organisation of dedicated workshops with interested services)	Regular meetings of the EU institutional stakeholders 1 to 3 workshops per year
Terminology community	Conferences (delivery of presentations)	1 to 3 conferences per year
Semantic Web community	Conferences (delivery of presentations: Multilingual Web conference, SEMIC…)	1 to 3 conferences per

2.3.10.3 Governance approach

The implementation of a governance body is needed at different levels.

In the first phase the work to be done in the scope of the ISA² programme needs to be governed by a representative subset of the stakeholders of the final system. This group is considered as the implementation of the initial government structure. It should rely on governance structure, which is being set-up for the CEF.AT platform (see http://www.lr-coordination.eu/anchor-points).

If the implementation of the system has been decided, the governance structure has to be adapted to be able to support and to supervise the management and the further evolution of real production system.

2.3.11 TECHNICAL APPROACH

Aspects to be considered:

- Management of multilingual taxonomies/terminologies

  The data structure for the public multilingual knowledge infrastructure will be defined by a core data model, which will be composed of two parts: a mandatory part (core metadata), which has to be respected by all data providers and optional part (private extensions) to allow the publication of additional data, which exist for a particular dataset and which are not covered by the core data model, if it represents an added value for the users of the system. I.e. additional data could be stored by the system but will eventually not be fully supported by the common services offered by the system.
  All individual concepts have to be represented in the adopted (semantic) format. Each individual concept will be identified by a unique persistent URI.
  The reuse/adoption of existing software components will be encouraged.

- Distributed infrastructure

  The public multilingual knowledge infrastructure should be implemented as a distributed network of RDF triple stores in order to guarantee a maximum of availability of the system.
  Adequate management capabilities are needed to guarantee the consistency of the data.
  APIs and online access should be implemented in a way that the technical implementation is hidden. The user works on a virtual system, which is composed a set of federated RDF triple stores, physically hosted in different locations.

- Management of the system

Also the management services of the system should be implemented in a way that the technical implementation is hidden.

There will be different groups of services:

- Maintenance of data
  - Ingestion of new data sets (including validation processes)
  - Update of existing data sets
    - Management of supplementary concepts, i.e. concepts that only exist on the level of the public multilingual infrastructure (create, update, delete)
    - Search and visualisation
- Maintenance of data structure (core data model and extensions, relations, alignments…)
- Management of the platform itself
  - Administration interface (monitoring of services, configuration, user management (for contributors and administrators)

2.3.12 COSTS AND MILESTONES

2.3.12.1 Breakdown of anticipated costs and related milestones

Phase: Inception Execution Operational	Description of milestones reached or to be reached	Anticipated Allocations (KEUR)	Budget line ISA/ others (specify)	Start date (QX/YYYY)	End date (QX/YYYY)
Inception	Project organisation has been set-up	60		Q2/2016	Q3/2016
Execution	Standard representation has been adopted	50		Q3/2016	Q4/2016
Execution	Core data model and a first set of extensions have been defined (including documentation)	100		Q4/2016	Q1/2017
Execution	Technical architecture has been defined	100		Q1/2017	Q2/2017
Execution	Proposal for an adequate government structure has been defined	50		Q1/2017	Q2/2017
Execution	First release of the system (operational proof of concept)	300		Q3/2017	Q2/2018
Execution	Proposal for the implementation strategy	60		Q1/2018	Q3/2018
Execution	Creation of the community	60		Q3/2017	Q4/2018
Execution	Feasibility study for the enhancement of the semantic capabilities of the platform	150		Q3/2016	Q2/2017
	Total	930			

2.3.12.2 Breakdown of ISA² funding per budget year

Budget Year	Phase	Anticipated allocations (in KEUR)	Executed budget (in KEUR)
2016	Inception/execution	160	
2017	Execution	445	
2018	Execution	325	

3. ACCESS TO DATA / DATA SHARING / OPEN DATA