This document is part of the Coordination and Support Action CRACKER. This project has received funding from the European Union's Horizon 2020 program for ICT through grant agreement no.: 645357.

Deliverable D3.4

Coordination with and Support of MLi

Authors: Khalid Choukri (ELDA), Vladimir Popescu (ELDA)

Dissemination Level: Public

Date: 03 August 2016

Status: Final

Grant agreement no.	645357
Project acronym	CRACKER
Project full title	Cracking the Language Barrier
Type of action	Coordination and Support Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	1 January 2015, 36 months
Dissemination level	Public
Contractual date of delivery	31/06/2016
Actual date of delivery	03/08/2016
Deliverable number	D3.4
Deliverable title	Coordination with and support of MLi
Type	Report
Status and version	Final
Number of pages	27
Contributing partners	ELDA
WP leader	ATHENA RC
Task leader	ELDA
Authors	Khalid Choukri (ELDA), Vladimir Popescu (ELDA)
Internal reviewers	Jan Hajic (CUNI), Stelios Piperidis (ATH)
EC project officer (M01-18)	Pierre-Paul Sondag (M01-M18), Susan Fraser (M19-36)
The partners in CRACKER are:	• Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany • Charles University in Prague (CUNI), Czech Republic • Evaluations and Language Resources Distribution Agency (ELDA), France • Fondazione Bruno Kessler (FBK), Italy • Athena Research and Innovation Center in Information, Communication and Knowledge Technologies (ATHENA RC), Greece • University of Edinburgh (UEDIN), UK • University of Sheffield (USFD), UK

For copies of reports, updates on project activities, and other CRACKER-related information, contact:

DFKI GmbH
CRACKER
Dr. Georg Rehm
Alt-Moabit 91c
D-10559 Berlin, Germany

georg.rehm@dfki.de
Phone: +49 (0)30 23895-1833
Fax: +49 (0)30 23895-1810

Copies of reports and other material can also be accessed via http://cracker-project.eu.

Contents

List of Abbreviations

UG(C)	User-Generated (Content)
EEA	European Economic Area
(S)MT	(Statistical) Machine Translation
LT(C)	Language Technology (Component)
NLP	Natural Language Processing
API	Application Programming Interface
DSI	Digital Services Infrastructure
REST	Representational State Transfer
SEO	Search Engine Optimization

1 Executive Summary

In this deliverable we document the efforts undertaken in the MLi FP-7 project, which are relevant to the CRACKER project. Namely, we will emphasize the achievements of MLi that are important to the basic goals of the CRACKER project itself, namely the definition and creation of a general framework for capitalizing on Machine Translation (MT)-oriented language resources.

The goal of MLi was to provide the foundations of a scalable platform for the joint development, enhancement and hosting of multilingual data sets, processing tools and basic services. Such a platform is expected to contribute to the development of cross-border EU e-commerce, currently hindered due to language and cultural barriers.

This deliverable reports on the results achieved by the MLi project, which are relevant to the CRACKER support action. Thus, the following points will be addressed:

- the MLi Hub, a reference architecture for making LT services available to the user community.
- the language resources, especially MT-oriented, that have been inventoried in MLi.
- the LTi cloud, prototypical instantiation of the MLi hub, centred on an LT providing service akin to the Japanese Language Grid, American LAPPS Grid (without the “constellation” aspects), or the PANACEA European initiative.
- the services in support of R&D: these included the survey of existing and emerging research disciplines and practices, in the context of continuing innovation in the commercial language technology sector, with focus on social media, big data and that subset of the cloud computing field with direct relevance to the future possibilities of a European language cloud (and/or an agency with its own data centers established in later stages of programmes such as CEF).
- the services in support of innovation: analysing the needs and requirements of language industry, enterprise sector, public sector and culture sector.
- outreach and exploitation: consultations with major players in the e-commerce arena have been undertaken.

The current report synthesizes the main achievements of the MLi project and is based, to a great extent, on the Second MLi Periodic Report (MLi Consortium, 2016).

2 Introduction[1]

MLi is a 30 months Support Action Project, under the domain of Content analytics and Language Technologies Developing plans and services, which finished on April 30th 2016.

MLi aimed at delivering the strategic vision, operational specifications and initial definitions and requirements sought for the European Multi-lingual data & services Infrastructure (MLi), formulating an actionable multiannual plan for its development and deployment, and fostering the multi-stakeholder alliances ensuring its long term sustainability.

To accomplish the MLi project outputs we have worked simultaneously on three dimensions: the technical aspects of the infrastructure, the managerial considerations of MLi, in terms of future operations, governance and sustainability, and the strategic aspects linked to the extension of the MLi infrastructure in the coming years. The components of the project are summarized in Figure 1.

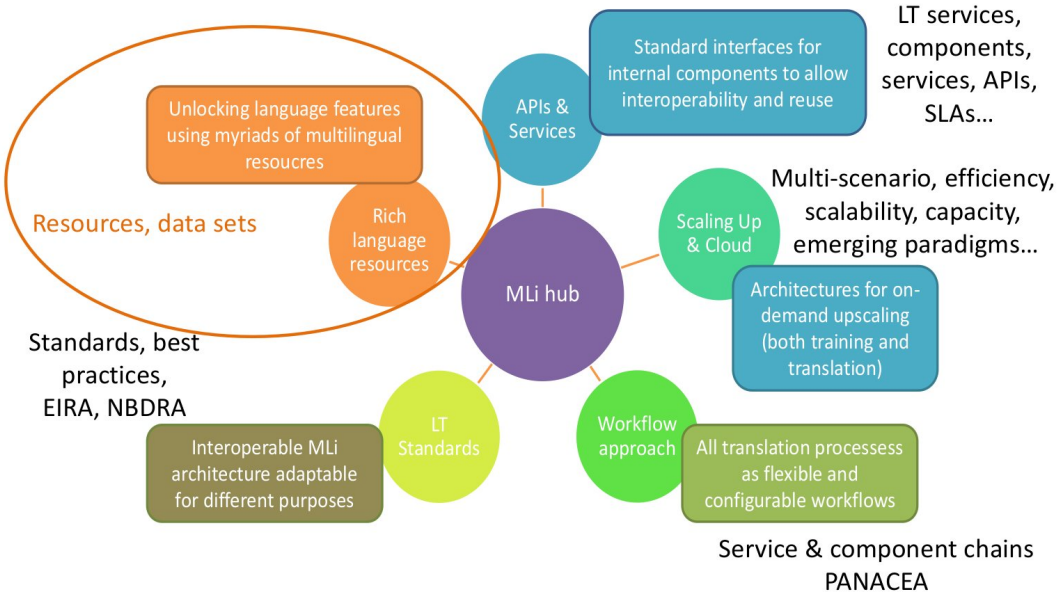

Figure 1 The MLi project roadmap

From the technical perspective, MLi has defined a layered architecture for the so-called MLi Hub (the proposed Multi-lingual data & services Infrastructure). This architecture is articulated in two axes:

- On the horizontal "Language Value Chain" the value is created by combining Language Technologies, Machine Translation and Natural Language Processing services and components into more complex and higher level features that fulfil requirements of the end user.
- On the vertical axis "IT Value Chain" the value is created by providing more abstract access to the system functionalities, starting from low-level infrastructural services, through LT/MT/NLP domain-specific services until the high-level

[1] This section is based, to a large extent, on (MLi Consortium, 2016: 5-6).

workflows and LT marketplace services for support of the business services and information and process integration within the Business Entity.

In this context, the LTi Cloud has been prototyped in the scope of MLi, as an instantiation of the MLI Hub architecture. It focusses and validates the upper layer of the platform, which focuses on the business and high-level service integration; it provides components of the MLi Hub such as the Broker, LT Marketplace and Access service facilities. The LTi Cloud prototype has been used as a tool to provide the market and LT stakeholders an “actual feel” of the architecture and how it will interact with them; as well as to refine the known requirements.

Thus, the LTi Could is a prototypical LTC Broker that provides a marketplace (or something similar to the App Store concept) that connects LT SaaS endpoints into one unified entry point for LT services. The main idea is to bring LT Vendors and LT Consumers together on one platform and overcome the decentralisation and fragmentation of Europe’s LT market. It is also a platform for rapid prototyping of LT systems that need to provide more sophisticated language stacks that combine various services from other providers.

The design of the Hub aims to be generic and use-case agnostic. It serves as a reference and a starting point for instantiating concrete LT oriented implementations, tailored to any concrete business case. However, in order to bring the design of the resulting MLi architecture closer to the real-word usage, the MLi project has provided applied and tailored sample scenarios to the concrete application to market areas, such as it has been the case of e-commerce.

From the managerial perspective, a comparative study of different modes and forms of governance and sustainability models that might be applicable for MLi future offerings has been performed during the project. A governance with an overview of Intellectual Property regimes and plausible business models has been produced based on the services provided by the LTi Cloud prototype.

The strategic perspective, acting as a bridge between recent and ongoing efforts in the LT field and the Data-Value Chain, MLi project has started fostering synergies with other initiatives related to the Europe 2020 strategy, namely the Connecting Europe Facility (CEF) programme along with its Digital Service Infrastructures (DSIs) as well as the efforts taken towards machine translation by the Directorate-General for Translation (DGT). In addition, MLi has produced a special report on e-commerce, as well as framed the MLi design into e-commerce use cases.

Further, the MLi project has delivered a tailored strategic vision and operational specifications for e-commerce in order to bring stakeholders closer to real-word usage. One of the main goals of MLi with respect to e-commerce has been to illustrate how to sell goods cross-border using LT tools to overcome language and cultural barriers. Even organizations that currently have cross-border and cross-language e-commerce platforms often do so by creating local branches that localize everything for the specific countries they operate in. Through MLi we have presented alternative solutions, where LT technologies might have a substantial role improving efficiency and cross-border market penetration. Among use cases that favour this approach, the MLi consortium tailored use cases with the use of multilingual search and SEO, MT or text analytics.

Overall, the MLi project has shown: a) the potential role of Language Technologies (LT) as an enabling component for growth and competitiveness in a Europe, which is currently moving towards a European Digital Single market; more specifically in the case of e-commerce; b) the potential of horizontal and vertical integration of LT with other solutions, or among LT solutions, which support the overcoming of the fragmentation of the LT supply value chain and the demand side. Moreover, the potential to move commercial and non-commercial organisations towards a European market with no language barriers, in which currently coexist multiple small market segments defined by the 60+ European languages.

3 MLi Hub[2]

The MLi project aimed at delivering "*the strategic vision and operational specifications for building the European MultiLingual data and service Infrastructure (MLi)".*

One of the major outcomes of the project is the technical and service platform design called the MLi Hub. From the e-commerce perspective, the MLi Hub aims at covering potential infrastructure needs for developing and packaging services and toolsets. Therefore, it is not a specific solution, but rather an umbrella that covers a set of LT activities, tools and services.

So far, after drafting a concise state of the art of the main initiatives related to the MLi Hub, a candidate reference architecture for the MLi Hub has been proposed taking into account several use cases and paying special attention to the fulfilment of e-commerce scenarios. The reference architecture describes the main building blocks that form the technical and service infrastructure. The architectural analysis has been made in the context of an iterative methodology and is based on the reference enterprise architecture comprising multiple aspects of the overall design of the platform.

In the MLi Hub architecture five elements are seen as fundamental:

- **Rich language features through the reuse of multilingual resources** – MLi Hub should allow for reusing as much as possible existing multilingual resources, such as those identified within the META-SHARE inventories. This implies taking into account various details such as resource metadata description, interoperability formats, resource storage, search, discovery and licencing.
- **APIs and Services** – natural language processing and machine translation comprise a plethora of smaller, specialized tasks, each one providing a different functionality. Those **services are typically chained together** in order to obtain a desired goal or provide a higher level service. Apart from language-based services, also a framework for third-party services, support services and even human-based tasks should be taken into account.
- **Scaling-up & Cloud-based deployments** – a very important aspect of the MLi architecture is the multi-scenario deployment capability. It must take into account requirements toward the stakeholders' business scenario, but also processing efficiency, data and resources capacity, and sustainability of computing infrastructure. Flexible architecture should take into account the ability of scaling up and scaling out, depending on the scenario. Also the cloud-based deployment should be considered where the computing resources can be acquired in the pay-as-you-go manner, giving the flexibility of dimensioning the overall solution to fit the increasing need for translation capacities.
- **Workflow approach** – machine translation systems typically perform training or translation as a complex process spanning various, often repeatable tasks. In a fixed scenario, such process is created up-front and does not change. Complex

[2] This section is based on (MLi Consortium, 2016), as well as on unpublished MLi reporting presentations.

workflows should be manageable so the system can reproduce the step-by-step process. This flexibility allows adapting a particular MT solution to different use cases or even languages, without the need of redesigning the whole solution.

– **LT Standards** – the building blocks of MLi Hub should comply with industry-established technology and service standards or best practices. This is crucial for ensuring architecture and component reusability, interoperability and adaptability. Relying on industry-agreed standards minimises the risk of technological obsolescence and enables the participation of technology suppliers, third party component developers and service providers.

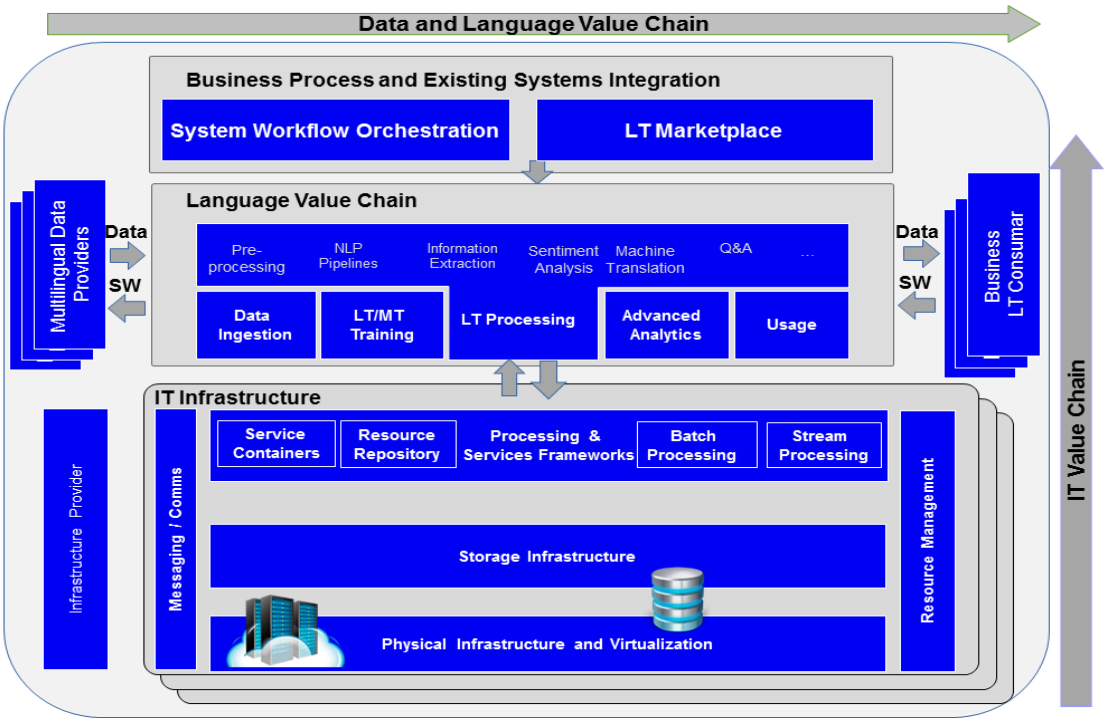

Figure 2. The MLi Hub Architecture

In the MLi Hub architecture depicted in Figure 2, several components participate together in observing the elements listed above:

– The **IT infrastructure** provides the technical and lower-level means for IT services delivery.

– The **Language Value Chain** is the central high-level building block encapsulating the whole process of LT/MT, from reading the input data to producing concrete information carrying a business value. It consists of specialized, domain-specific and scenario-specific technical components and data resources enabling LT/MT services.

– The **Business Process and Existing Systems integration** encapsulates operational details and functional capabilities of the devised LT/MT architecture. It contains a system workflow orchestrator, which combines technical Language

Value Chain services into a concrete business workflow. It also contains a service marketplace, which aggregates platforms to offer LT/MT to third parties.

In the following paragraphs, we will provide further details on these components, while also providing an exhaustive listing of all the components.

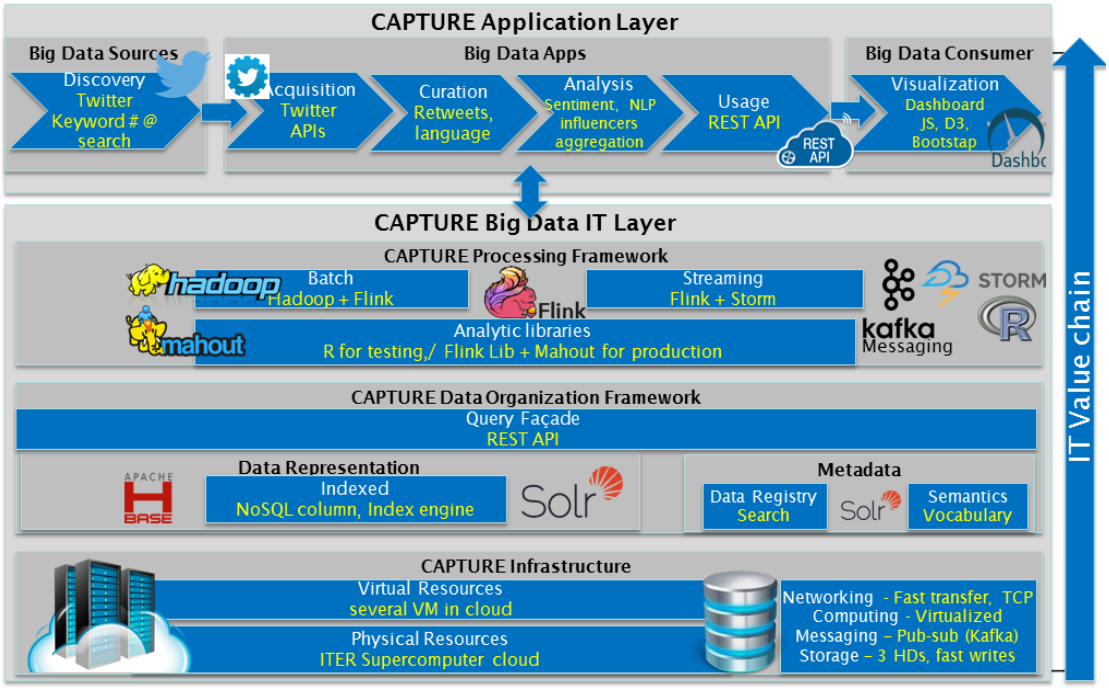

Figure 3. Sample implementation of MLi Hub reference architecture. Social Networks monitoring service, "Capture"

ATOS has implemented the MLi Hub reference architecture in a real-world example. The example is a Social Networks monitoring service, called "Capture". As shown in Figure 3, on the Physical Infrastructure and Virtualization layer, Capture uses specialized cloud infrastructure and storage facilities. For the Storage layer, Capture proposes a set of NoSQL databases and indexed repositories to cope with the requirements posed by the streaming nature and high data volumes associated to social networks, along with a query façade to ease the access to the storage. On top of that, Capture proposes the use of several streaming and batch processing tools and frameworks (Apache Hadoop, Apache Flink, etc.) in order to serve as foundation of the analytical process needed by the LT components in a big data environment.

The Language Value Chain in "Capture" is performed in the Application Layer. This layer implements the Language Value Chain by providing a set of steps and analytics using LT tools and services. The results of the process (such as vectors with quantitative measures) are delivered via RESTful services to the customers (e.g., in the JSON format) and visualized (e.g., as a graph) using a web-based dashboard.

The main building blocks of the MLi Hub Reference Architecture depicted in Figure 2, reading the figure bottom-up and left-right, are the following:

IT Infrastructure – This block provides technical and lower-level support for IT services delivery and technological provisioning for such areas as: IT physical infrastructure, infrastructure virtualisation (in case of the cloud deployment

scenarios), storage infrastructure (either centralized or distributed), IT storage and processing services, such as resource management, IT service provisioning infrastructure and processing infrastructure. It is important to mention that in certain scenarios this block may be implemented in very different manners. For instance, the infrastructure might be in-house (a single instantiation of the application in a single infrastructure, distributed or not); or outsourced to a single third party infrastructure provider (in case of adopting IaaS or other cloud-based solutions); or covered by several third-parties infrastructures (in case of the Hub providing third-party services hosted onsite in the providers infrastructures). In this sense there is no “one-fits-all” IT Architecture scenario, but rather a tailored solution to the particular business use case.

Physical Infrastructure and virtualization – consists of low-level IT stack, such as servers and network infrastructure, solutions for dimensioning IT resources for certain capacity requirements, operating systems virtualisation and management infrastructure, cloud infrastructure and isolated environment management.

Storage Infrastructure – provides means for data storage in a potentially distributed infrastructure. In case of large-scale deployment, this might include distributed file systems, highly-scalable distributed database Infrastructure, redundant, failsafe and replicated storage systems, etc.

Processing and Service Frameworks – based on the lower level IT stack, this IT service Infrastructure delivers services for abstract access to the IT Infrastructure, regardless of the physical and organisational distribution and configuration.

Resource Repository and Management – services for storage and accessing of concrete data items (resources) within the system. Examples of the resources hosted can be large corpus of language-pairs for machine translation, sentiment data, etc. One examples of a systems that could be used to implement this is Meta-Share.

Service Containers – a technical backbone for the service framework and service marketplace, providing a container and isolation for flexible service deployment and scaling. Example of this is the service model transformations needed in a brokerage system such as the one provided by the LTi Cloud.

Batch and Stream Processing Infrastructure – access to the computational resources for executing computational tasks on the underlying infrastructure. Examples of these processing engines can be Map-Reduce frameworks, stream processing software components, etc.), but in general it covers any processing infrastructure needed to run the services and the analysis performed in the upper layers of the architecture.

Infrastructure Providers – general term for third party IaaS and PaaS providers who facilitate the IT Service infrastructure as a whole or a particular part of it. While instantiating the MLi Hub architecture might require certain IT Architecture effort, the physical part might be in-house or acquired as-a-service from a third party if the business and technical requirements are satisfied. In the current version of the LTi Cloud, the infrastructure for running the LTi Cloud services is minimal, but the Language Value Chain services are actually hosted in the infrastructure of the providers.

Language Data Providers – data assets, central to the particular Business Entity that are subject to LT/MT processes in order to provide an added value business

service. E.g., product descriptions in the e-commerce company to be translated into different languages in order to provide a cross-border business service.

Language Value Chain (LVC) – the central high-level building block encapsulating the whole process of LT/MT, from reading the input data to producing concrete information that is used to provide a business value. It consists of specialized, domain-specific and scenario-specific technical components that facilitate and implement concrete LT/MT services. They are mostly business-driven, functional components that provide system-wide means to support and realize business objectives.

Data Ingestion – accessing and reading input documents either from existing systems (e.g., company's CRM system, CMS, Social Media, etc.) or from specialized user interfaces. E.g., eCommerce site, user-uploaded documents or texts to translate.

LT/MT Training – one of the core activities within the language technologies required is to provide a custom, domain-specific trained models for concrete LT/MT task. Note that the training is not an integral part of the LVC, as many systems would use only translation services already trained if needed. However, as one of the core activities of MLi, it is considered as an integral part of the LVC. The training consist of LT Resources (multilingual resources for training that might be stored and accessed via the Resource Repository), and training algorithms and models (domain-specific algorithms for model training, typically particularized to some concrete aspect of LT/MT/NLP task. An example of this is MOSES.

LT Processing – technical components for performing concrete LT/MT/NLP services. Those might be small low-level NLP services (such as, text pre-processing, language detectors, segmenters, stemmers, parsers, etc.) or higher-level and more complex ones (e.g., named entity recognition, statistical translation, sentiment analysis or question answering systems). These services can be deployed in a single or several IT Infrastructure, as for instance in the case of the LTi Cloud, where similar services can be selected from different service providers. It is not coincidence that this layer stays on top of other LVC components, due to the fact that it typically needs other components, such as trained models, data ingestion or advanced text analytics in order to provide a concrete service. A comprehensive study of components and APIs has been performed in several MLi work packages.

Advanced Analytics – technical component providing higher level features and services for advanced analytics, not necessarily focused only on text analytics, such as Artificial Intelligence systems, Business Intelligence systems, large-scale media mining, etc.

Data Access Usage – a set of software APIs for accessing processed data, such as translated documents, processed texts, extracted features or advanced search results. It provides data and service consumption interfaces to the Business Entity.

Business Process and Existing Systems Integration – this building block encapsulates operational details and functional capabilities of the devised LT/MT architecture. It is the top-level service layer, and contains most notably:

System workflow orchestrator – component that combines technical Language Value Chain services into a concrete business workflow that might also include other actors, specific o the business context of the Business Entity. The objective is to

realize defined business goals in a continuous manner. This component could be implemented using existing tools and frameworks (e.g., PANACEA) or use simpler mechanisms based on data flows, pipelining or messaging.

Broker – this component is needed when the system provides functionalities to mediate between different LT services offering similar characteristics. It is in many cases essential when the intention is to offer a Service Marketplace, but it is probably not implemented when the platform is intended for specific applications that use a known set of services. An example of brokerage is the LTi Cloud prototype.

Service Marketplace – it is an umbrella term for service aggregate platforms to offer LT/MT services to third parties. This block may not be necessary for simple LT installations, while may scale to a fully-fledged service marketplace offering a menu card of services including pricing and delivery mechanism. An example is the LTi Cloud prototype, which main objective is to offer a LT services marketplace.

Business LT Consumer – this building block represents the final user of the architecture. The final user is the ultimate LT consumer that uses the results of the LT Value Chain for their own business objectives (i.e., MT of an e-commerce site, social media analytics, a developer using services of a LT Services Marketplace – such as the LTi Cloud – etc.).

4 Language Resources

One of the goals of the project is **to define what language resources are needed in terms of data and software/tools to populate and enrich the MLi platform**. These resources represent the requirements to carry out (applied) research, (pre-) commercial technology development as well as service deployment, in and between EU languages. Both the needs of developers and deployers are taken into account if, for instance, the latter carry out some system adaptation themselves.

The different types of LRs required to perform MT with a focus on e-commerce have been identified: bilingual User Generated (UG) corpora (to train the system and produce a translation model), monolingual UG corpora (to produce a language model that will smooth the translation output), and lexical resources, ideally containing non-standard expressions that allow to normalize the input to be further processed.

An inventory of available bilingual and monolingual resources for the 24 EU languages (and for the corresponding 276 language pairs) has been elaborated on the basis of data available in the OPUS platform[3]. The identified data are available in a format exploitable by the Moses system, a statistical machine translation (SMT) system allowing to train translation models for any language pair. The number of aligned sentences available per language and per language pair is provided, thus allowing to draw a general picture of the existing resources and of the languages for which further resources will need to be produced.

The histogram depicted in Figure 4 makes it evident that, even if all 24 EU languages have some aligned data with all the other languages, the volume of aligned data for some of them is considerably lower. Maltese, Croatian and Irish, for instance, have between 1.6M and 2M aligned sentences, whereas other languages, namely German, Swedish, Italian, Danish, Portuguese, Dutch, Spanish and French, have over 45M aligned sentences. **The robustness of the MT systems that can be built for each language may thus not be homogeneous.**

[3] OPUS is one of the largest collections of translated texts from the web. It consists of freely-available online data, harvested, aligned, enriched with linguistic annotations, and provided to the community as a publicly available parallel corpus (open content package).

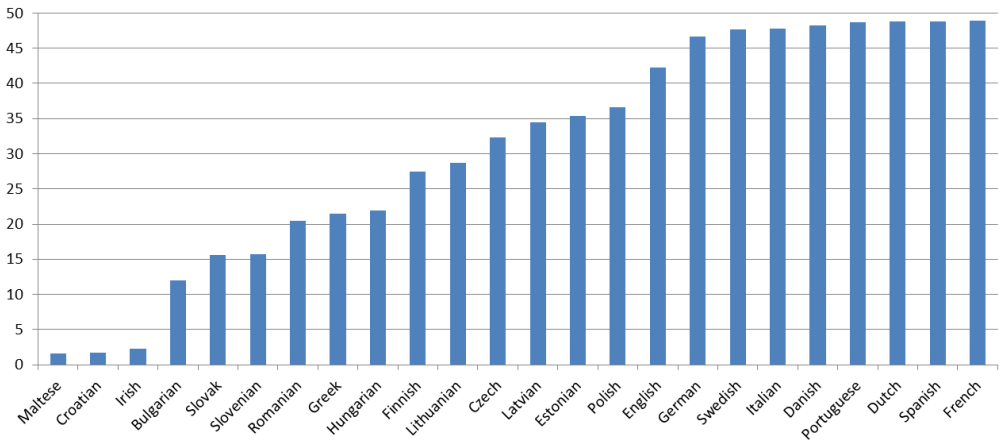

Figure 4. Number of parallel sentences per language (in millions). Source: Own elaboration on the basis of individual OPUS corpus statistics

In terms of **data volume**, MT results show that increasing the amount of data is of help to improve MT quality, but **this improvement tends to flatten as the data sizes become very large**. For highly-inflected and agglutinative languages, even large amounts of data do not guarantee adequate coverage of all the forms and phenomena. For these languages morphological analysis (PoS taggers) may be more effective than a large data volume.

From the research carried out so far in the project, it has also become clear that **LRs specifically involving UGC are scarce**. Bilingual UG corpora are the hardest to obtain. Nevertheless, the web, and particularly social media (such as Twitter), is a rich source of user-generated data that can be compiled with the aid of available tools. Some of these tools have been identified and described, such as monolingual and bilingual web crawlers produced in the framework of previous EU projects (e.g., PANACEA).

Data sparsity was identified as an issue to take into account in MT applied to e-commerce. Given the great variety of topics, genres and registers produced by users it is challenging to create MT systems which are generally adequate to any e-commerce domain. Therefore in several situations novel LRs will actually have to be produced, e.g., an opinion corpus is not necessarily adequate for training an automatic translator targeting rather impersonal and standardized user reviews, given the different linguistic features characterizing each textual genre. In Figure 5 we summarize the types of translatable contents in the e-commerce context; we notice that UGC occupies an important place, especially, in the dynamic realm.

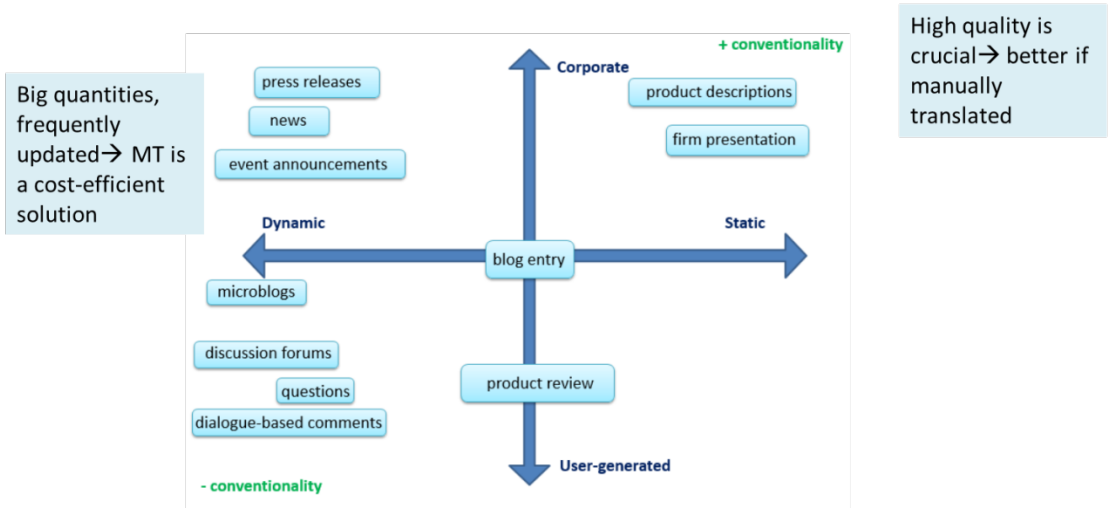

Figure 5. Types of translatable content in e-commerce web sites

With regard to the scarcity of UGC aligned data to train domain-specific MT systems, a contribution of MLi has been to point to available solutions tackling this issue:

- **Pre-editing** has been suggested as a part of domain adaptation in the translation of forum posts (Jachmann *et al.* 2014). Pre-editing is aimed in this context at bringing UGC closer to standard text.
- The SMT system can also be refined by including a pre-processing step in which **potential spelling errors are modelled** (e.g., through a Confusion Network) and subsequently recovered by the decoder on the basis of a character *n*-gram language model (Bertoldi *et al.* 2010).
- Adaptation through **pseudo-in-domain data**: another option is to adapt MT to a given text type by using parallel data from a generic domain that resembles that of the specific domain, so-called 'pseudo-in-domain'. In the case of UGC one could consider as pseudo-in domain data datasets such as subtitles, etc. (e.g., the OpenSubtitles corpus).
- **Monolingual MT adaptation**: another possibility is to use in-domain data for the construction of the language model. For example one could crawl vast amounts of tweets with Twitter API. For 'small' languages this would be far from ideal, but specific tools are available.
- **Bootstraping UGC MT systems with the aid of crowdsourced translations**: active learning techniques could be used to select, from an in-domain pool of UGC data, the minimum subset that would produce the best performing system, e.g., selecting data that is relevant and reducing redundancy among the selected texts. Next, one could translate this data in the target language by means of crowdsourcing (e.g., Jiang *et al.* 2012). With regard to crowdsourcing, however, additional costs in terms of quality check (e.g., through BLEU scores) should be added. This 2-step approach should thus be thought of as an investment, which might make sense for a company depending on the expected ROI.

The results obtained so far set the ground for a more detailed analysis relying on specific case studies and datasets to assess the reuse potential of out-of-domain data (e.g., non-UG data, or UG data from other domains) for developing MT systems targeting UGC produced in e-commerce environments. The costs of production, evaluation, exploitation and deployment of newly created resources have also been assessed (see Table 1). These considerations have lead to a final set of recommendations defining a LR roadmap for multilingual EU e-commerce (Choukri et al. 2016).

Thus, an inventory of available bilingual and monolingual resources for the 24 EU languages (and for the corresponding 276 language pairs) has been elaborated on the basis of OPUS corpora. The corpora that have been taken into account are: Europarl, EUConst (EU Constitution), Acquis communautaire, OPUS KDEdoc, OPUS KDE4, OPUS Open Office, OPUS PHP. These corpora were also taken as a point of reference in the Euromatrix project, which aimed at promoting MT for all pairs of languages of the EU.

The number of aligned sentences available per language and per language pair is provided, thus allowing to draw a general picture of the existing resources and of the languages for which further resources will need to be produced.

Table 1. E-commerce MT-enabling strategies

	Bilingual corpus (translation model)	**Monolingual corpus (language model)** **Price for IPR clearing**	**Total cost (per language pair)**	**Total cost 25 language pairs (CEF languages)**
Scenario 1: From scratch	200.000 € (0.5M words)	5 000 €	205 000 €	5 125 000
Scenario 2: Adapting TM and LM	15 000 € (25K words)	5 000 €	20 000 €	500 000 €
Scenario 3: Adapting LM	0	5 000 €	5 000 €	125 000 €
Scenario 4: Using general MT systems	0	0	0	0

Therefore, an action plan has been prepared, which addresses work in LR identification, production and an appealing sharing roadmap which encourages cooperation. Thus MLi can help focusing community efforts towards resource sharing via inventories like META-SHARE and CLARIN/LINDAT. On the other hand the availability of metadata (e.g., licensing conditions) and data from these inventories can help foster the adoption of cloud-based solutions such as the LTI Cloud.

This roadmap assumes several scenarios, going from the most expensive one consisting in building MT technologies into e-commerce sites from scratch, to using general, off-the-shelf MT systems, such as those provided by widely-available search engines. They are all depicted in Table 1.

Certainly, scenario 1 is the most generous and, potentially, technically the most successful one. However, given its prohibitive costs, it is unlikely that the bulk of SMEs is able to afford it. Hence, Scenarios 2 and 3, based on “mere” adaptations of the translation models (TM) and/or language models (LM) seem much more appropriate. Scenario 4 is the “poor man”’s approach to leveraging the multi-lingual capabilities of e-commerce platforms and is only appropriate for providing rough translations of UGC, like reviews or comments.

5 LTi Cloud

The MLi team used the analogy of the app-store to describe the LTi Cloud. This way, the LTi Cloud prototype is described as an app-store like web directory for Language Technology Components. Its main goal is to bring together consumers and providers of Language Technology. It targets companies building software solutions with a stack of sophisticated language processing capabilities and who want to evaluate the fit of a concrete Language Technology Component.

The LTi Cloud allows the consumer to discover LT components and to run free trials, whereas the LT vendors can prepare and promote LT components. It is also a business platform that allows marketers to advertise and feature their company's LT components.

The LTi Cloud responds to the industry needs identified in a survey amongst the leading European Language Technology companies. The survey, to which more than 200 companies were directly invited to participate, revealed that companies are struggling with high development costs related to the provision of broad language support. For many companies extending their language coverage has been reported to be cost-prohibitive (Hummel and Kranias 2014).

Thus, as shown in Figure 6, the LTi cloud bridges the gap between industrial users (SMEs, DSIs, IT Integrators, etc.) and language technologies that can leverage the linguistic capabilities of the applications driven by these industrial entities. In turn, the language technologies disseminated through the LTi cloud are delivered either by research actors (e.g., universities, public or private research institutions, etc.), or by the industry itself (e.g., companies whose core business revolves around language technologies).

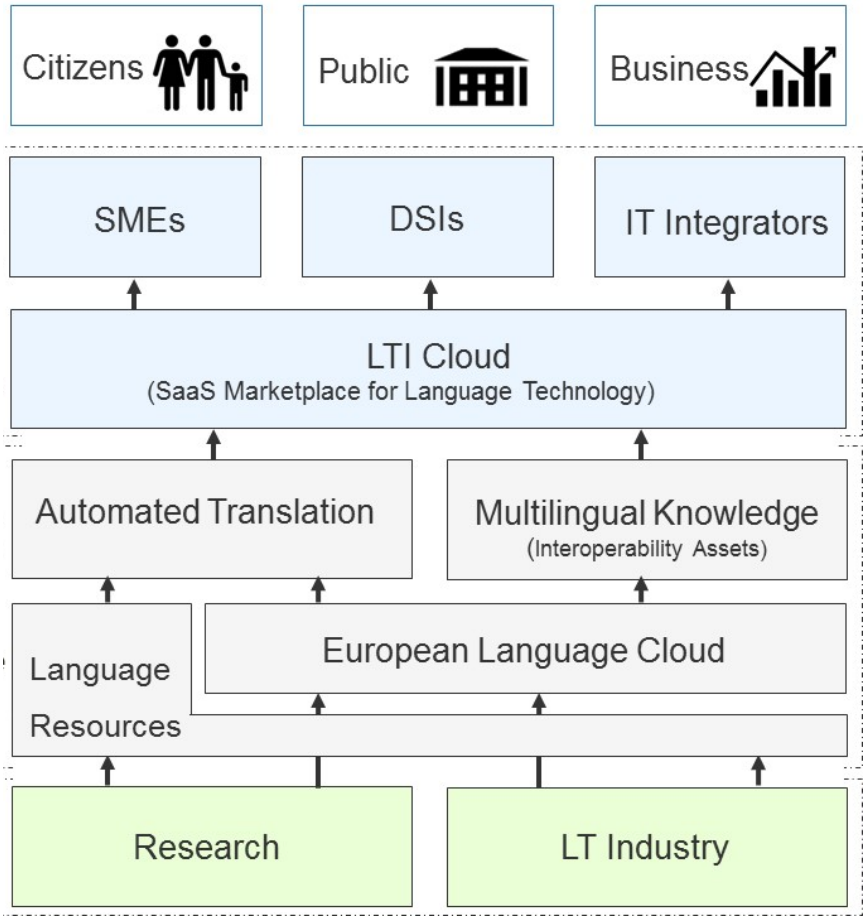

Figure 6. Place of the LTi cloud in the EEA Digital Single Market landscape

5.1 LT Solutions for E-commerce

One of the main goals of MLi with respect to e-commerce is to help organizations to sell goods cross-border using LT tools. In practical terms this means the potential increase of selling from a reduced market to all EU countries and beyond. Cross border EU e-commerce market is indeed expected to grow in the next years. Whereas in 2014 cross-border e-commerce was about 8% of total EU e-sales, in 2020 it is expected to reach about 18%.

However, so far even organizations that currently have cross-border and cross-language e-commerce platforms often do so by creating local branches that localize everything for the specific countries they operate in. This is a sub-optimal solution that LT technologies may have a substantial role to improve. Technologies such as improved multilingual search and SEO, MT, text analytics, speech recognition, etc., are key to favour this approach.

More concretely, MT has been identified as the potential solution to one of the challenges currently hindering e-commerce, namely, the language barrier. The availability of a pan European platform enhancing the development of multilingual technologies would be an asset for enterprises aiming to enter e-commerce. Indeed, such a platform would lower dramatically the costs for developing MT services, thus

accelerating ROI and encouraging SMEs with limited resources to develop multilingual e-commerce sites.

In this respect, several existing technologies can be used to leverage the multilingual capabilities of e-commerce sites. Thus, cloud-based service providers offering access points to remote LT components are suited to the task.

Some of these allow the users to define processing pipelines and workflows. Others propose off-the-shelf complex components.

Panacea[4]: targeted at NLP Practitioners: allows to explicitly define workflows; there are actually several web services proposed.

The Japanese Language Grid[5]: Targeted at a wider audience: allows its users to explicitly define web services associated to language resources, thus exposing APIs that users can tap on, in order to build runnable applications (e.g., for Machine Translation).

The LAPPS (Language Application) Grid[6]: Targeted, as PANACEA, at NLP and CL practitioners, this grid allows its users to set up Web Services and build workflows aggregating such services, in order to have applications corresponding to classical NLP tasks at the end, e.g., POS tagging, Dependency Parsing, etc.

5.2 The LTi Cloud Produced in the MLi Project

In MLi, a language services cloud has been produced, as a prototypical **app-store** for Language Technology components (LTCs), which brings together **LT providers** and **consumers**, and chains various LTCs through standard APIs within the LTI-Cloud proxy. The architecture of the resulting **LTI cloud** is depicted in Figure 7.

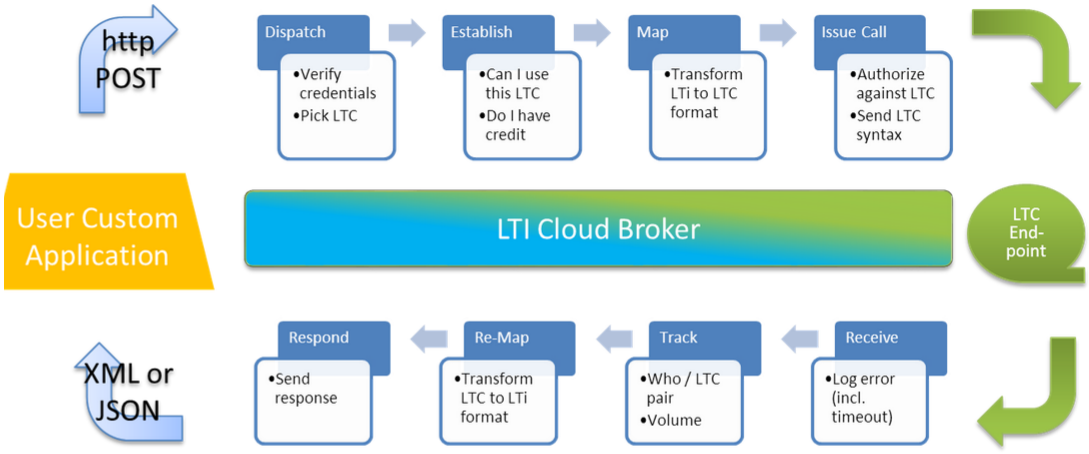

Figure 7. The LTI Cloud Architecture

[4] http://www.panacea-lr.eu/

[5] http://langrid.org/en/index.html

[6] http://www.lappsgrid.org/

The LTI Cloud acts like a proxy and maps requests and responses from a standard LTI-Cloud format to the proprietary LT Vendor formats. It has been developed by Esteam[7] and is available from: https://lticloud.eu/.

In order to meet the challenge of providing a complete set of NLP-related services, an Application Programming Interface (API) has been specified and developed. This API is supposed to give the ability to developers to take advantage of Natural Language Processing (NLP) tools in their applications in a way that no domain or internal knowledge will be required.

For each vendor-provided processing tool a specific component (LTI Broker) is developed in the LTI Cloud which handles the data conversion between the vendor-specific format and the LTI Cloud format. Thus, technical requirements providers should meet have not been explicitly defined as the LTI Broker is customised for each vendor-provided tool (Wetzel et al. 2016: 14).

The consumption of the proposed linguistic API targets a huge number of possible clients. In this regard, a set of principles to fulfil requirements coming from the different type of users were defined (Hummel *et al.* 2015), including:

- Open and modular service oriented architecture based on REST that can easily support new content sources, new language services
- Wrapping of language services
- Decoupling of the NLP API and the consumer application
- Usage of common data formats (JSON, XML) for data transfer in order to ensure interoperability between different architectures. As the API is exposed as a service layer, a number of commonly used formats should be used in order to comply with recent best practices in data exchange protocols. JSON is an excellent way to provide a universal protocol for the Language Enabler API data exchange. It is simple and fast to parse but at the same time powerful enough to represent complex data structures.The XML format is also supported for the Language Enabler API.
- Domain agnostic interface for re-use purposes
- Support for various backend linguistic tools in order to give the ability to use the appropriate method given a specific problem.

The business model for the LTI Cloud has also been defined, including a roadmap, a sustainability plan, and potential organizational structures (Hummel *et al.* 2016). For this purpose, the characteristics and interests of the three actors targeted by the LTI Cloud (Language Technology consumers, the Language Technology providers, and the operators of the platform itself) were analysed. Also, the interests and roles of the different players in this market place for language technology were balanced and evaluated.

[7] https://www.esteam.se/.

6 Services in Support of R&D and Innovation

The activities performed in this task comprised the analysis of requirements and provision of recommendations related to Big Data in the following areas:[8]

- Platforms/Infrastructure: this comprised the assessment on the persistent experiment repositories for language science such as the ones used by the EC in different translation cases or in the infrastructure set-up, components and tools for MT optimization and deployment, and XaaS market places
- Language Technology: the assessment comprised the optimization of MT systems for user-generated content within big data, designing specific MT systems for blogs, forum posts, etc.

The activities comprised the assessment of requirements and provision of recommendations related to User Generated Content (UGC) in the following areas:

- Principled strategies, which include the prioritization of translation of user-generated content from the strategic perspective.
- Platforms/Infrastructure: this comprised the assessment on pool access to tweet streams and archives, surface web fragments and their partitions, and computational resources sufficient for their processing, analysis on the device frontier: social/mobile/embedded, identification of the parameter set for successful deployments processing social media graphs and streams, etc.
- Content: this comprised the assessment and recommendations for scalable infrastructural support for linked data interoperability.
- Language Technology: activities comprised assessment and recommendations for the optimization of MT systems for user-generated content within big data, designing specific MT systems for blogs, forum posts, etc., as well as for tailoring MT systems to individual social media content types and language-specific features.

The activities performed comprised to identify, characterise and lay the foundations of a number of high-impact innovation driven public-interest services and innovation services geared towards the digital single market. The activities focused on deep analysis and specification of key services that are important for the language industry, enterprise sector, public sector (eGovernment), security and Digital Single Market, and facilitates crossing the language barriers. The analysis and specification of essential service categories demonstrated, that the language technologies, especially automatic translation services, are important and requested by different sectors.

For this purpose, work done involved identifying and reviewing the potential needs and current experience on end-user systems and plug-in components for (a) trans-European public interest services (mainly but not exclusively in the public sector) and (b) cross-national/sectoral services geared towards Europe's digital single market.

Work performed combined activities for analysing MT services for public sector, as well as making eGovernment and commercial services interoperable and enabling knowledge based data processing through the Multilingual Knowledge Cloud

[8] This section is mostly based on (MLi Consortium, 2016).

approach. This included the analysis to extend and elaborate on semantic interoperability and knowledge assets to cover the needs of eGovernment services.

More specifically, the assessment was performed on how LT can cover data and coverage gaps be filled, the opportunities for eGovernment regarding security and monitoring, healthcare (ageing, assisted living, public health information, prevention, cross-border prescriptions and emergency patient records access, early warnings based on social networks analysis), pan-European business analytics, eGovernment and eParticipation (communication with citizens, including cross-border), cross-border communication, mobility, e-commerce and customer care and services, logistics, eLearning and gaming across Europe.

The activities also comprised the presentation of positive examples of MT usage in public sector, such as the translation service for EU presidency in Latvia and machine translation service hugo.lv developed by Latvian government, or the need for NATO Strategic Communication Centre of Excellence to monitor news and social media content which can be predictive of and help detect hostile, and or terroristic cyber-attacks and counter the impact of the use of the internet for terrorism. Further, the Machine translation for Connecting Europe Facility is analysed.

Work included the analysis of the usage of the language technology services in Localisation and Globalization industry and assessed the multilingual challenges and innovation work for different market sectors. The assessment included: a) the financial sector, where example solutions that could address the end-user needs in multilingual and actionable information were analysed; b) the e-commerce sector where analysis was carried out for two major e-commerce providers; c) the public sector in Europe where efforts were focused on evaluating the MT services for public sector (also the impact of language barriers to information exchange in legislation was analysed); d) the need for technological toolkit of automated translation, textual pattern and keyword tracking, and text analytic functionality for cyber security and defence in the public sector.

7 Conclusions

This deliverable detailed the efforts undertaken in task 3.2: “Coordination with and support of MLi”. The MLi project, initially aimed at providing the key insights, strategic vision and architectural underpinnings of a pan-european multilingual data services infrastructure, for fostering the multilingual capabilities of EEA (European Economic Area) public and private actors, narrowed down its focus to the European e-commerce field, in the context of the European Digital Single Market (DSI). The main objective has become to provide the necessary ingredients to make it so that in the EEA someone who is working with e-commerce should not have to bother with the data formats and workflows that go on inside the machine. The goal is built-in and transparent interoperability.

The MLi project partners reported to the CRACKER Consortium regularly and also during the project meetings. Some of the debates that took place within CRACKER were also shared with the MLi team whenever this was considered as relevant.

8 References

Bertoldi, N., Cettolo, M., and Federico, M. (2010) Statistical machine translation of texts with misspelled words. In *Proceedings HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 412-419.

Choukri, K., Fernández-Barrera, M., Popescu, V., Gaspari, F., Toral, A., Way, A. (2016) D3.3 – WP3 Final Report. Comprehensive specifications to be input to the drafting of future calls with a detailed Language Resource roadmap. MLi Deliverable WP3-D3.3.

Hummel, J., and Kranias, L. (2014) Language Enabler Requirement Analysis Report, ESTeam AB. MLi Deliverable WP4-D4.1. Available at http://mli-project.eu/wp-content/uploads/2014/11/MLi-D4.1-Requirement-Analysis-Report-26-nov-v02.pdf

Hummel, J., Kranias, L., Magnusdottir, G., Wetzel, M. (2015) Specification of the Language Enabler API. MLi Deliverable WP4-D4.2.

Hummel, J., Wetzel, M., Kranias, L., Magnusdottir, G. (2016) LTI Cloud Business Model Report. MLi Deliverable WP4-D4.3.

Jachmann, T., Grabowski, R., and Kudo, M. (2014) “Machine-translating English forum posts to Japanese: On pre-editing rules as part of domain adaptation”. *Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing* (ANLP), (pp. 808-811). Sapporo, Japan, January.

Jiang, J., Way, A., Ng, N., Haque, R., Dillinger, M. and Lu, J. (2012) “Monolingual Data Optimisation for Bootstrapping SMT Engines”. *Proceedings of MONOMT 2012: AMTA 2012 Workshop on Monolingual Machine Translation*, San Diego, CA.

MLi Consortium, *MLi Project Periodic Report*, no. 2, Spring 2016; see also the MLi deliverables, accessible at http://mli-project.eu/?page_id=490.

Wetzel, M., Hummel, J., Kranias, L., Magnusdottir, G. (2016) LTI Cloud Prototype. MLi Deliverable D4.5.