European Platform for the Multilingual Digital Single Market

To unleash its full potential, the Digital Single Market (DSM) must be multilingual. Language technologies are the key to creating a truly multilingual DSM and crossing language barriers in Europe. Drawing on years of innovative research and development, much of it supported by EU funding, European industry has developed multiple mature technologies to enable multilingualism.

At the same time industry offerings are very fragmented as companies either focus on larger languages only or on their local markets. Many EU languages are disadvantaged due to gaps in technology coverage or poor quality compared against a handful of larger languages, for which much better solutions are available. As a result many European companies and citizens are basically excluded from future key innovations.

If the fragmented industry cannot deliver multilingual capabilities for all member states, the public sector should make a difference by kick-starting the creation of the necessary infrastructure. Investing in such infrastructure will yield an incredible ROI by enabling a truly cross-lingual DSM. This initiative must go hand-in-hand with efforts to address other obstacles and barriers to establishing the DSM. Only by removing all these restrictions together, Europe will be able to fully achieve its goals of a DSM.

In order to create a truly multilingual DSM, the language technology industry proposes to create a **European Platform for the Multilingual DSM (Multilingual Platform)**. This Multilingual Platform will combine mature language technologies (LT) in several clouds, and provide services which will be made available to startups, SMEs, IT integrators, industry, and the public sector.

These capabilities are essential to enable SMEs and integrators to create multilingual solutions that will cross language barriers. These solutions also will be fully integrated into the workflow of EU institutions and national Member States.

The Platform will encompass three layers: solutions (Layer 1); infrastructures (Layer 2); and research (Layer 3) corresponding to the Strategic Agenda for Multilingual Digital Single Market document (Strategic Agenda).

Innovative Solutions (Layer I)

Layer 1 combines language technology solutions built by SMEs using the services available in the Multilingual Platform. Solutions will enable e-commerce and digital services providers, public administrations, cross-border public service providers, and other stakeholders to easily integrate multilingual capabilities in their daily work. These solutions can also be integrated into the workflows of large-scale organizations – such as EU institutions, NGOs, media outlets, and corporations – as well as be made available to the public, enabling multilingualism on a pan-European scale (see Strategic Agenda for more details).

Enabled by the basic language technology services of the infrastructure clouds (Layer II), LT SMEs and solution providers will create components to cover a range of client and market specific needs. Fragmentation of these commercial offerings will be overcome by the establishment of the LTI Cloud - the LT industry's one-stop- marketplace for LT components.

LTI Cloud – the Marketplace for Language Solutions

In order to make it easier for solution providers to discover, test, integrate, and license Language Technologies, commercial LT components will be aggregated in the LTI Cloud. The LTI Cloud is a SaaS wrapper around language technology (LT) components and functions as a marketplace. It will make it easy for start-ups, IT departments, system integrators, and software companies to plug 'n' play language technology. LT companies can market their capabilities, build LT components based on others, and use the LTI Cloud as a customer acquisition channel.

Language Service Infrastructure (Layer II)

The Infrastructure layer will combine the mature language technologies in four distinct clouds: *Automated Translation Cloud, Human-Computer Interaction (HCI) Cloud, Multilingual Knowledge Cloud,* and *European Language Cloud*. The clouds are interoperable and based on a core foundation of language resources. These infrastructures must not compete with the software industry or service providers. They provide fundamental facilities and systems which private companies cannot build efficiently or profitably.

The governance of the infrastructure layer should involve a broad coalition of stakeholders. We propose the creation of the **Coalition for Multilingual Europe** (Coalition), made up of representatives from industry associations (e.g., LT-Innovate, GALA), research organizations (META-NET, CLARIN, etc.), European institutions (DG Connect, DG Translation, etc.), and key user groups (e.g. BDVA). The Coalition will serve as the main partner of the Commission. The Coalition will be tasked with overseeing and coordinating the operations of the infrastructure components, ensuring a competitive, open, and transparent implementation process.

Automated Translation Cloud

Combines the highest-quality machine translation services for each EU language and all language pairs, provided by CEF.AT, EU Member States, and commercial providers. These MT services will be available in multiple domains. The services will enable solution developers to integrate instant translation capabilities into any platform or application, including mobile apps, web portals, or e-commerce sites (see Solutions Layer I).

CEF.AT services: MT services provided by CEF.AT for the CEF DSIs and EU public administrations.

National MT services: MT services developed by EU Members States for use in national public sectors (e.g., Hugo.lv, Versti.eu).

Commercial services: MT services developed by commercial providers (e.g., LT-Innovate and TAUS members).

Human-Computer Interaction (HCI) Cloud

Proliferation of ubiquitous devices like tablets and mobile phones, and increasingly also smart appliances and robots, requires efficient human-computer interaction beyond traditional graphical and text based interfaces. Voice commands are quick and intuitive, allowing to control devices in a natural way. Though the latest research has yielded strong results in voice based interaction (e.g. Siri and Cortana) this success is confined to larger languages. Speakers of smaller languages do not have access to technology that would allow them to interact with devices in their own language.

The HCI Cloud will address this issue by providing speech and other human-computer interaction services for all EU languages, which can then be used to build robust multilingual solutions.

The HCI Cloud combines speech services – automatic speech recognition (ASR), text-to-speech (TTS), dialog management (DM), speaker and language identification, keyword spotting, voice search, audio and video indexing using phonetic and word level indexes, dysfluency detection and removal, specific tools for connecting automatic translation and speech processing for spoken text translation tools, and multimedia communication modules for basic functionality in sign language recognition and generation, image search, image and video object recognition and tracking, text-in-image and -video recognition – for all EU languages.

The Human-Computer Interaction Cloud will be based on mostly Open Source tools, such as the KALDI speech recognition generator based on state-of-the-art machine learning methods, which will in turn utilize the language resources collected and annotated. Similar technology exists for the other tasks, such as text-in-image identification, and object detection in images and video.

Multilingual Knowledge Cloud

The Multilingual Knowledge Cloud combines semantic interoperability services for making eGov and commercial services interoperable and enabling knowledge based data processing. It is essential for the implementation of the European Interoperability Framework (EIF) which is planned to become mandatory for all public IT projects. Semantic Interoperability required by the EIF is achieved by deploying multilingual meaning and knowledge assets. These assets will be pooled and exposed in the Multilingual Knowledge Cloud. It will enable meanings to be carried across language boundaries via data structures and data elements that are specific to different sectors. The Multilingual Knowledge Cloud will embrace existing developments at EU institutions. For example, the European Office for Harmonization created the multilingual knowledge system TMClass in order to be able to process Community Trade Mark applications in all official languages.

ISA (Interoperability Solutions for European Public Administrations) can provide the standards and processing know-how for aggregating Semantic Interoperability Assets, making them discoverable in a central network and promoting their usage, as well as filling-up of missing languages. It would define and host the access layer so that larger IT companies, SMEs, and start-ups will be able to build on these assets and flourish by combining them with other types of data to build and market new applications.

European Language Cloud (ELC)

All language processing applications (search, mining, writing, speech, translation, etc.) depend on a basic natural language processing (NLP) infrastructure. The European Language Cloud (ELC) is a public infrastructure which provides the basic functionality required to process unstructured content. Through an API it provides basic language technology services such as tokenization, stemming, morphology analysis, part of speech tagging, named entity detection, identification of measurements, currencies, formulas, etc. for all languages, in the same base quality, under the same favorable terms. National institutions in charge of language maintenance provide data and standards. The ELC builds on language resources and forms the basis for all LT efforts in text and speech processing.

Language Resources

At the core of the various clouds making up the platform is a repository of language resources. These resources include language data such as parallel and monolingual texts, lexicons and terminologies used for building MT services, voice recordings used to build ASR services, and monolingual texts used for developing linguistic tools. Also included are the language resources provided

by CEF.AT. The resources will be provided by the European language research community, as well as by various EU Member States through the efforts of pan-European language-data collection initiatives.

Research (Layer III)

Layer 3 addresses the gaps in coverage for all EU languages, and provides novel methods to improve quality and applicability of language technologies. This Layer is described in details in the Strategic Agenda.

Diagram of the European Platform for the Multilingual DSM

The chart below illustrates the above described platform for enabling a Multilingual Digital Market. In layer I the IT industry provides innovative solutions for Europe's citizens, business, and public administration. Layer II provides the necessary infrastructure for Europe's IT companies and it makes sure that all EU languages are supported. In Layer III Europe's researchers in academia and companies are inventing and developing the technology for tomorrow's solutions.