Challenge 4: Technologies for Digital Content and Languages

Digital content remains the material basis for a multilingual knowledge based society. However, the explosive growth of digital content (both structured and unstructured) makes it important for European citizens and organisations to learn to manage it effectively and to extract from it maximum value in terms of private or public, personal or organisational decision making, planning and management,.

The focus of Challenge 4 is on:

- exploring and testing new approaches, methods and techniques to extract, interpret and exploit information from unstructured multilingual and/or multimedia sources, yielding actionable knowledge;

- developing and testing in realistic operating conditions new algorithms and software frameworks to analyse and visualize extremely large volumes of data in real time;

- supporting Small and Medium Enterprises (SMEs) developing innovative applications in structured and unstructured digital content management and, particularly, in the reuse of open data.

Support actions for road-mapping are also envisaged to prepare the constituencies for bringing together research and innovation aspects in Horizon 2020.

Objective ICT-2011.4.1 Content analytics and language technologies

Target Outcomes

Due to the combined effect of globalisation and European integration, there is a growing need for effective solutions that support multilingual business and inter-personal communication, and enable people to make sense of digital services in Europe's many languages.

a) Cross-media content analytics

Innovative methods for mining unstructured information embedded in text, speech, audio and video for the purposes of context-aware interpretation, correlation, aggregation and summarisation, turning information into usable understanding and actionable knowledge. Special emphasis is placed on social and collective intelligence from multilingual sources. Projects shall achieve broad coverage with efficient semantic interpretation. Of specific interest is the ability to capture sentiment and represent concepts and events, identify relations and similarities, interpreting time and space, within and across individual media, thus increasing our ability to detect and exploit otherwise hidden meaning across a range of applications.

b) High-quality machine translation

Advancing machine translation (MT) by pushing the research frontier and bridging relevant disciplines. Emphasis is placed on high-performance and easily configurable MT yielding high-quality translations suitable for publication with little or no human intervention. Expected innovations include effective hybridization of existing and emerging solutions, the ability to autonomously learn from use and human feedback, and to adapt to new situations with high portability and scalability. Work should cope with everyday language and with the need to compile translation resources dynamically from the web or enterprise repositories. Projects are expected to demonstrate the successful integration of MT within larger systems.

c) Natural spoken and multimodal interaction

Speech-enabled interfaces based upon multimodal verbal and non-verbal communication. Projects shall address autonomous human-like social agents that can handle conversational speech; learn from interaction and react proactively to new communicative situations; recognize and generate social cues. Systems should be able to cope with spontaneous dialogue and exhibit adequate communicative, affective and cognitive (e.g. question answering) capabilities in relation to the domain/task under consideration and the needs and abilities of the user. Technologies should be designed to match multiple delivery platforms, from virtual assistants e.g. for customer service, through smartphones to games.

For each of the target outcomes (a), (b) and (c), the call invites

(i) a few ambitious R&D projects investigating new approaches and research avenues well beyond the current state of the art; projects will be centred on cross-disciplinary approaches and partnerships, and address multimedia content and multimodal interaction; they will encompass everyday language as found in e.g. consumer-generated content, cover multiple languages and cater for written and/or spoken language as appropriate; technologies shall be adaptive, cope with massive volumes of content, and have a clear potential to support real-life processes;

(ii) one broad-based support action designed to establish a unifying roadmap in each of the domains under consideration, developing a compelling research and innovation agenda until and beyond 2020, centred on close collaboration between research centres and commercial players (particularly SMEs), and based upon agreed reference architectures, common resources (standards, software, data), and shared development and evaluation facilities.

d) Developing joint plans and services

The call invites one support action intended to design and lay the foundations of a scalable platform for the joint development/enhancement and hosting of (multi-)language data sets, processing tools and basic services. The action will build upon and extend existing and emerging collaborative infrastructures. The aim is to create over time a comprehensive online repository of reusable modules and components, in the broadest possible range of EU languages, underpinning research, technology transfer and industrial development efforts.

Expected Impact

- Strong participation of private-sector players, including SMEs, well above the FP7 ICT average.
- A unifying research roadmap aggregating the vision of more than 200 centres; a common innovation agenda based on the business strategy of more than 100 companies.
- Technological leadership and increased innovation capacity as a result of widely accepted roadmaps encompassing presently fragmented communities.
- A European open-source MT system becomes the most widely adopted worldwide; post-edited MT becomes the standard mode of translation within 5 years, increasing significantly ($> 25\%$) the efficiency of human translation.

Funding Schemes

a), b), c): STREP, CSA

d): CSA

Indicative budget distribution

- STREP: EUR 21 million

Call:

FP7-ICT-2012-10

Objective ICT-2011.4.2 Scalable data analytics

Target Outcomes

– Tools and skills to deploy and manage robust and highly performance data analytics processes over extremely large amounts of data. User-driven research with public and methodologically sound quantitative performance evaluation criteria is a strict requirement. As a bridge to activities to be pursued under the Horizon 2020 program, two distinct types inter-disciplinary road-mapping activities can be supported: a roadmap for networking and hardware optimisation in support of next generation Big Data management solutions and a second roadmap for the social, legal, economic study of externalities in the (re)use and linking of data.

a) Scalable algorithms, software frameworks, visualization

- Novel algorithms, software infrastructures and methodologies for real time interaction, visualization, analytics and decision support applications over extremely large volumes of data (both structured and unstructured).

- Data types that are currently experiencing very high growth rates are of special interest including (but not limited to) 3D, biology, genomics, financial, geospatial, social networks, transportation, logistics, telecommunications, engineering, and any type of data stream.

- Non-traditional database and storage solutions and data integrity protection tools are solicited for the robust integration of heterogeneous data sources such as static and streaming data.

The availability of extremely large and realistically complex data sets and/or streams is a strict requirement for participation as is the availability of appropriate populations of experimental subjects for human factors testing in the domain of usability and effectiveness. Software implementations must be rigorously tested in the environment of professional organisations with a clear stake in the solution and a clear path to deploying it, if effective.

b) Big Data networking and hardware optimisations roadmap

- One inter-disciplinary CSA to bring hardware and networking experts together with designers of algorithms and software frameworks and Big Data practitioners. This will define a shared European vision for future Horizon 2020 R&D activities on the design of dedicated processing or networking hardware for optimising the performance of Big Data analytics, including programming frameworks that software developers without specialised hardware knowledge could use easily.

The roadmap will chart advances in scalability and run-time performance as well as energy efficiency and sound methods for analysing and optimising capital versus operating costs of Big Data operations. The CSA will also be responsible for

disseminating the roadmap across relevant constituencies and establishing cross-disciplinary communities with a shared understanding of concrete problems worth investigating in future programmes.

c) Societal externalities of Big Data roadmap

- One CSA to produce a roadmap for future activities on the sharing and reuse of large and linked datasets including those obtained by data harvesting across heterogeneous data sources. The CSA should bring together social science scholars, open data activists, statisticians, computer scientists and other relevant parties in order to design a European data environment capable of amplifying positive externalities and reducing negative externalities.
- Positive externalities to be addressed include (but are not limited to) economic and legal models for efficient data markets.
- Negative externalities include (but are not limited to) the privacy risks that come from the re-identification of personal information, particularly as a consequence of more and more data sets becoming available and being linked to one another. Ethical and moral considerations should be taken into account.
- The CSA will also be responsible for disseminating the content of the roadmap across the relevant constituencies and establishing cross-disciplinary communities with a shared understanding of concrete problems worth investigating in future programmes.

Expected Impact

- Advanced querying and analytics applications with sub-second response times over distributed information resources consisting of trillions of records.
- Ability to query or detect in real time complex events against dynamic feeds of millions of data streams generating hundreds of thousands of events per seconds.
- Visualization systems enabling exploratory analysis and manipulation without any perceptible delay on data resources containing billions of items.
- Enabling European suppliers to reach by 2020 a share of the Big Data market compatible with the size of our economy (30% of world market).

Funding Schemes

a): IP, STREP

b), c): CSA

Indicative budget distribution

- IP/STREP: EUR 26 million
- CSA: EUR 5 million

Call:

FP7-ICT-2013-11

Objective ICT-2011.4.3 SME initiative on analytics

Target Outcomes

Helping European Small and Medium Enterprises acquire the competence and resources they need to develop innovative content and data analytics services. Development of services based on the reuse of available data, particularly from public bodies, is specifically required for theme a) and encouraged for theme c).

a) Integrated Open Data Reuse Incubator

An Integrated Project to establish an environment and calling for efficient, small scale development of services of commercial interest based on the use of European open data by Small and Medium Enterprises (SMEs). The IP should:

- devote most of its resources to publish and manage regularly scheduled and well-advertised calls for SMEs to submit mini-proposals to be funded for a period between six and twelve months.

- create a computing infrastructure where the winning mini-proposals will find accurate, up-to-date and (when useful and feasible) linked versions of the data they need for their services and, if they so wish, deploy the experimental version of their services.

- establish a mechanism for connecting open data demand and supply by systematically contacting European public bodies about their open data availability and assisting them in the efficient and sustainable publication of such data, if needed with targeted engagements.

- solicit open data reuse ideas from the general public and conduct a European wide open data reuse information campaign.

- The IP will finally create a process to connect the most successful SMEs with sources of funding and business networks.

b) Easing transfer and take-up of language technologies

Language technologies are often deployed within products and services relating to web or enterprise intelligence, including text and audio mining, social media analytics and sentiment analysis, enterprise search and content management, online and cloud based translation, etc.

This action targets focused user- and market-oriented projects in any of the above areas, with the overall goal of bringing language technologies closer to commercial maturity through an "industrialisation" process including but not limited to: i) engineering of promising but commercially untried technologies, e.g. in terms of performance, robustness and coverage; ii) integration within existing or upcoming products and services; iii) first-use experimentation and validation in a clearly identified application domain; iv) in-depth assessment along technical, used related and economic dimensions; (v) identification of possible exploitation paths and viable business models, and of suitable sources of funding.

c) Software components and intuitive end user applications based on reuse of open data

Development of software components supporting the whole lifecycle of reuse of multilingual open data, particularly from public bodies. This includes

i) usable data publications methodologies and tools, adapted to the operating conditions of typical public bodies and rigorously tested for traceability, usability and sustainability in a public body environment;

ii) methods and tools for linking open data sets produced by public bodies;

iii) methods and tools for optimising open data applications based on public demand both in terms of content and in terms of functionalities/usability;

iv) cross platform development tools for delivering intuitive and responsive open data applications on multimodal devices and environments such as mobile, tablets as well as desktop.

Expected Impact

- A European open-source MT system becomes the most widely adopted worldwide; post-edited MT becomes the standard mode of translation within 5 years, increasing significantly (>25%) the efficiency of human translation.
- Dozens of data application software components, used by hundreds of developers.
- Hundreds of applications, reusing billions of open data records, used by millions of end users around the EU
- Wider creation of valuable applications by integrating available public data with the users' own data, including contextual information available from mobile devices.

Funding Schemes

a): IP

b), c): STREP

Indicative budget distribution

- IP: EUR 5 million,
- STREP: EUR 15 million

Call:

FP7-ICT-2012-SME-DCA