



**Roberto Cencioni**

**European Commission**  
**DG Information Society & Media**

*Digital Content & Cognitive Systems*  
*Language Technologies*

*Brussels, 19-21 June 2012*



- we promote and support **multinational** research & innovation projects
  - focussing on the digital world, both economy & society
- our aim is to teach computers how to process – if not to understand - **written & spoken human language**
  - if you master human language, then you can try & cope with **multiple languages**
  - and there are plenty...
    - 60+ languages in Europe and 23 official working languages in the EU





- the **European Union** brings together **500 million people** speaking **many languages...**
  - half of the EU adult population don't speak English
  - no such thing as a *digital lingua franca* since the advent of the social Web, smartphones, e-readers...

## **EU policy framework: "Digital Agenda for Europe"**

- ✓ build a ***digital single market*** where content & services can flow freely
  - ✓ support cross-border exchanges between public online services
    - ✓ ease internationalisation of SMEs...

*but what if you don't understand what's being displayed?*





## EC has supported Language Technologies since 1980s

- a fresh start since 2008
  - renewed political commitment
  - explosion of online content, esp. social networks
  - sizeable industry of language services
  - promising technical advances
- approx **150 M funding** in the last 5 years
- nearly **60 projects** underway
- next *call for proposals* in July





## main features of our projects

- multi-party, multi-nation partnerships
  - member states, associated countries, 3<sup>rd</sup> countries
- academia + industry + "launching" users
- focus on enabling technologies
- multiple languages
- impact
  - target groups, application scenarios, demonstrators
- significant involvement of commercial players





## main themes of our projects

*"... effective solutions that support business and inter-personal communication and enable people to make sense of online content and services in Europe's many languages."*

- global content processing, eg automated translation
- mining of unstructured information, eg text analytics
- natural interaction, eg speech understanding & dialogue

in and between multiple languages



# Our approach



we struggle since 2009 to

- boost cross-national & -disciplinary **partnerships**
  - from 27 project proposals in 2009 ...
  - ... to 200 proposals in 2011
- address **gaps & weaknesses**
  - in particular duplication & fragmentation
    - common research **vision**
    - **reuse** of basic assets (data & tools)
    - technology **transfer & deployment**





- **research:** towards a federating vision & roadmap
  - META-NET (2010-2012)
- **business:** ease market uptake & deployment
  - LT-INNOVATE (2012-2014)
  - forum for business stakeholders esp. vendors
  - also, market data, value chains, non-technical barriers, demand & supply
- **public sector:** promote the use of LT by our own institution
  - inter-service group with major EC departments







## **completing the current programme (2012-13)**

- last series of calls in July, 4 topics:
  - 1) (cross-media) content analytics
  - 2) (high-quality) automated translation
  - 3) (speech-enabled) multimodal interaction
  - 4) industrialisation of "intelligence" related LT technologies

## **launching the new programmes (2014-2020)**

- a. Horizon 2020 (H2020), research & innovation*
- b. Connecting Europe Facility (CEF), pan-European service platforms*
  - concrete orientations by Q1 2013





## **last series of calls** under the current programme:

- Challenge 4 – *Digital Content & Languages* of the ICT work programme 2012-2013, released in July
  - Call-10: open Jul 12; close Jan 13; budget: 27 M (language)
  - SME call: same schedule; budget: 15 M (data + language)
- **Call-10**, R&D, objective 4.1:
  - a) cross-media **content analytics**
  - b) high-quality **machine translation**
  - c) speech-enabled **multimodal interaction**
- **SME-Call**, innovation, objective 4.3b:
  - b) **industrialisation** of “intelligence” LT technologies



## for each **Call-10 line:**

- endeavours which are ambitious, up- & out-ward looking
  - multi-disciplinary, multi-lingual
  - multi-media, multi-modal
- **1 road-mapping action**  
to help define, organize & bootstrap future collaborative work under H2020
- **1 to 3 projects**  
acting as a bridge towards H2020  
... social, mobile, cloud, big & smart data ...





## **SME call: 4.3.b Easing transfer and take-up of LT**

**... bringing language technologies closer to commercial maturity through an "industrialisation" process:**

- (i) *engineering* of promising but commercially untested technologies
- (ii) *integration* within existing or upcoming products and services;
- (iii) *first-use experimentation* and validation;
- (iv) *in-depth assessment* along technical, user related and economic dimensions;
- (v) *identification* of possible exploitation paths, viable business models, suitable sources of funding.





## **new programmes, 2014+**

- research (H2020)
  - innovation (H2020)
  - infrastructure (CEF)





## ***H2020 2014-2020:***

- **research & innovation;** budget (EC prop): 80 B,  
~20% for ICT
  - science (including FET)
  - *technologies* (most of the mainstream ICT themes)
  - societal challenges (e.g. health, ageing, education...)
- **on-going discussions with Council & EP**
  - outcome depends on concomitant discussion of *MFF*  
(= money for all EU policies)
- **meanwhile, early priorities & operations by Q1 2013**
  - first series of calls towards Q4 2013





## ***CEF 2014-2020:***

- **TEN model:** pan-European, public-interest infrastructures
- 3 pillars: Transport, Energy & ICT (new, 9 B proposed)
- for ICT: (i) Broadband & (ii) ***Digital Service Platforms***
  - eGovernment services, eHealth, eCulture...
  - **Multilingual Access to online Services** (new)
- aim is to "enable service providers to offer their content & services in as many languages as possible..."
  - core platform (hub) + general-purpose tools & data sets + services co-developed with industry & government
- still a long way to go vis-à-vis Member States, update in the autumn





## ***CEF Multilingual:***

"... enable providers of online services to offer their content & services in as many EU languages as possible, in a most cost effective way."

addressing:

- **public services** esp. eGovernment, eHealth, eCulture, Open Data portals
- **commercial services** eg eCommerce
- where possible, **user-generated content** (social media)







- less than 40% of Europeans (outside UK, IE) understand English, less than 10% *accept to use* online services in English
- multilingual services are necessary to:
  - unify the EU online market – from 20+ “sub-markets” to one single market of 500+ million consumers
  - enable the provision of public services across national borders
  - unleash the power of global online communities





## **"Core platform"** (75-100% EU funding)

### ***service infrastructure:***

hw/sw facilities; technical standards & APIs; service & legal provisions; user registration & help-desk; developers' accreditation & support...

### ***basic resources:***

aggregation, harmonisation & curation of 3<sup>rd</sup> parties' language data & tools, to be made freely (?) available as *building blocks* for the **development** of generic & specific services

### ***provision of services:***

hosting and/or pass-through platform for the **operation** of generic services





## "**Generic services**" (75% EU funding)

**co-developed** services enabling administrations & businesses to deliver their own integrated solution

- "horizontal" (sort of middleware), in support of other CEF domains
- "vertical" i.e. as proposed by industry & member states

in areas such as

- automated translation, (to *understand* content)
- multilingual information access & web mining (to *find* content)
- content management solutions (to *embed* content & create value-add services)
- ...





## Conclusions:

- heavy competition with other policy areas
- emphasis on competitiveness, growth & jobs
- so to be successful:
  - clear & credible **R & I agenda**
  - well understood **impact** on economy & society
  - well connected & mobilized **community**
  - broad (incl national) **recognition & support**





## Upcoming calls for proposals

### *ICT Work programme 2012-2013*

- Challenge 4 “Technologies for Digital Content & Languages”
- 2 objectives, 2 calls both launched in July:

**Call 10 - 4.1:**                    open July 2012  
   close Jan 2013  
   budget: 27 M  
   - dedicated to language

**SME Call - 4.3:**                    same schedule  
   budget: 20 M  
   - both content (data) & language

- approx. 12 LT submissions retained for funding





## Call 10 - 4.1 Content analytics & language technologies

- **new title** to reflect a broader, multi-disciplinary approach
- **continuity** in terms of project lines:
  - *content analytics*  
aka information extraction, mining of unstructured content, categorisation & summarisation, social media mining & sentiment analysis, business intelligence...
  - *machine translation*
  - *spoken & multimodal interaction*
- **discontinuity** in terms of ambition & timeframe  
- paving the way for work under H2020 (2014+)





- **as with previous calls, a few common features:**
  - written and/or spoken language, as required
  - multi-lingual (i.e. multiple in/out languages), where relevant cross-lingual (“translation”)
  - handle everyday language, social media & UGC
  - cope with massive volumes & diverse sources
  - cater for contextualisation & personalisation
  - technologies are generic & adaptive (language, domain, task)
    - but testing within specific application environments
- **new emphasis on**
  - *go beyond pure text or speech:*
    - 1) **multi-media** content & **multi-modal** interaction, and therefore
    - 2) **multi-disciplinary approaches & partnerships**
  - *progress across languages & media*





### 3 research lines (“outcomes”):

- a. (cross-media) content analytics
  - b. (high-quality) machine translation
  - c. (natural) spoken & multimodal interaction
- no cross-over, stay within the line you’ve chosen
  - **for each line** there is room for:
    - **1-3 projects** (75% funding) investigating new approaches & research avenues beyond the current SoA
    - no more than **1 support action** (100%) laying the foundations of coordinated efforts under H2020 (eg project clusters)

each line provides opportunities for

- **ambitious** forward-looking efforts, as well as
- active co-operation with **progressive** vendors & users





## projects: a) analytics

- some features:
  - exploring, interpreting, 'understanding' ... rich, unstructured content: interplay of text, speech, audio & video
  - exploiting diverse, multilingual sources
  - delivering significantly better-than-SoA accuracy across languages & media mixes (EN text being the benchmark)
  - coupling accuracy with efficiency & broad coverage
  - based upon linguistics, semantics, statistics & whatever else it takes...!
- what we don't need:
  - just "*yet another project*" with no significant S&T innovation





## projects: b) translation

- what we don't need:
  - projects delivering no significant improvement over today's MT quality
  - a mere re-combination of existing methods & techniques
  - another MOSES-based project with no substantial innovation
  - things others can do: post-edited MT, CAT & MT, crowd sourcing...

## projects: c) interaction

- what we don't need:
  - research with no clear *grounding* in existing or emerging ICT platforms
  - speech recognition, transcription, translation, acoustics... without
  - ... a prominent *dialogue* component
  - ... the interplay of speech & other *modalities*





## **roadmapping actions** (*blueprint for future R&I*)

- *sectorial* (see a-b-c), not for the LT community as a whole
- centred on needs, scenarios, tasks... defined & agreed with industry
- based upon common tools, data sets & standards; agreed reference architecture(s)
- joint integration & evaluation facilities ...
- ... yielding credible plans, efforts & timescales

## **budget breakdown**

- projects (STREP, 75%): no less than 21 M
- roadmapping actions (CSA, 100%): no more than 6 M
- no predefined allocation between the 3 project lines





## as with previous calls:

- single entry-point (Mr Kimmo Rossi)
- help-desk for inquiries & outline proposals
- info-days ...
  - *further details in due time*
- candidate experts/evaluators are welcome





## SME Call - 4.3 SME initiative on analytics

- **3 project lines:**
  - a. Open Data Reuse Incubator
  - b. Easing Transfer and Take-up of LT
  - c. Software and Applications based upon Reuse of Open Data
- **key parameters** (different from Y2011 call):
  - up to 2 years, up to 1.5 M funding
  - 2+ SMEs, > 30% of the funding
  - 1-step evaluation, 20 pages maximum





## 4.3.b Easing transfer and take-up of LT

... bringing ["intelligence" related] language **technologies** closer to **commercial maturity** through an **industrialisation process** including:

- (i) *engineering* of promising but commercially untested technologies
- (ii) *integration* within existing or upcoming products and services;
- (iii) *first-use experimentation* and validation;
- (iv) *in-depth assessment* along technical, user related and economic dimensions;
- (v) *identification* of possible exploitation paths, viable business models, suitable sources of funding.





## what's an SME?

- an **enterprise** which has
  - fewer than 250 employees
  - an annual turnover not exceeding 50 M
  - or an annual balance-sheet total not exceeding 43 M
- **relationships** with other enterprises must be taken into account (notably independence)
- the **official definition** of SMEs can be found at

[http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index\\_en.htm](http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index_en.htm)



***Thank you!***

**FP7 ICT programme:**

[http://cordis.europa.eu/fp7/ict/programme/home\\_en.html](http://cordis.europa.eu/fp7/ict/programme/home_en.html)

**Horizon 2020:**

[http://ec.europa.eu/research/horizon2020/index\\_en.cfm](http://ec.europa.eu/research/horizon2020/index_en.cfm)

