META-NET

META-NET: Language Technology for Europe

Georg Rehm

Network Manager META-NET DFKI, Germany georg.rehm@dfki.de

ČESAR META-NET Roadshow Event – Bratislava, Slovakia June 07, 2012



Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).

Outline



- Introduction
- META-SHARE
- Language White Paper Series
- Strategic Research Agenda

Multilingual Europe



- Challenge: Providing each language community with the most advanced technologies for communication and information so that maintaining their mother tongue does not turn into a disadvantage.
- While research has made considerable progress in recent years, the pace of progress is not fast enough to meet the challenge within the next 10-20 years.
- All stakeholders researchers, LT user and provider industries, language communities, funding programmes, policy makers – should <u>team up for a major dedicated push</u>.





META-NET is a network of excellence dedicated to fostering the technological foundations of the European multilingual information society.

META-VISION: Building a community with a shared vision and strategic research agenda

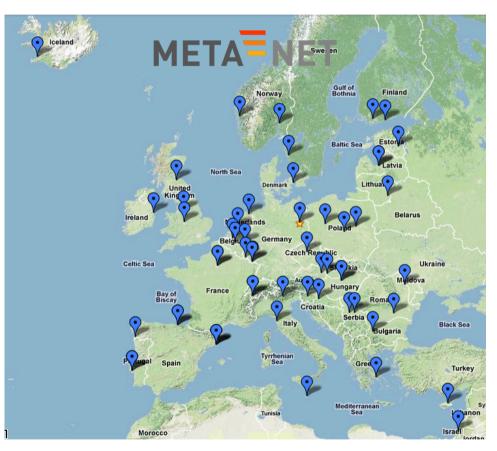
META-SHARE: Building an open resource exchange infrastructure

META-RESEARCH: Building bridges to neighbouring technology fields

Four Funded Projects



- Initial project: T4ME (FP7; 13 partners, 10 countries)
- Three new consortia (ICT-PSP) since Feb. 2011: CESAR, METANET4U, META-NORD
- All EU member states and several non-member states covered.
- META-NET in June 2012: **57** members in **33** countries.



http://www.meta-net.eu/members



META-SHARE

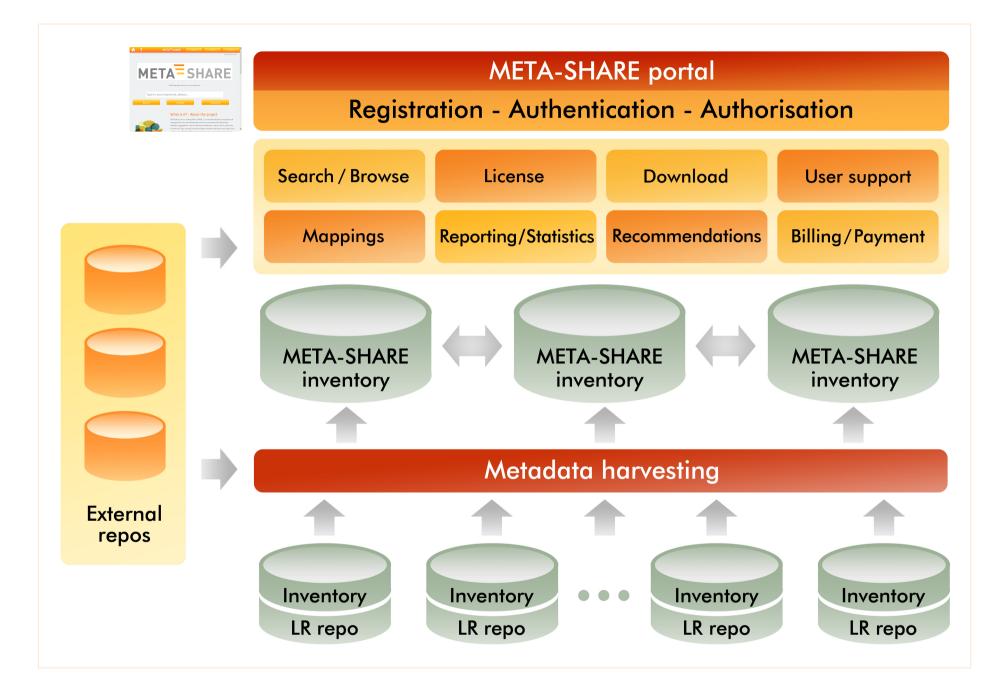
META-SHARE

http://www.meta-net.eu

META-SHARE at a Glance



- Open exchange infrastructure for language resources and tools.
- Language resources and tools are documented, uploaded, stored in repositories, catalogued, can be downloaded, shared, discussed.
- Improve their visibility, documentation, identification, availability, preservation, interoperability.
- Long-term goal: boost research, technology and innovation through wide availability, pooling, openness and sharing of resources.
- Repositories store and maintain inventories of resources and tools.
- Metadata inventories are exported and harvested in the network.
- □ Currently 13 repositories up and running; ~1.300 LRs available.

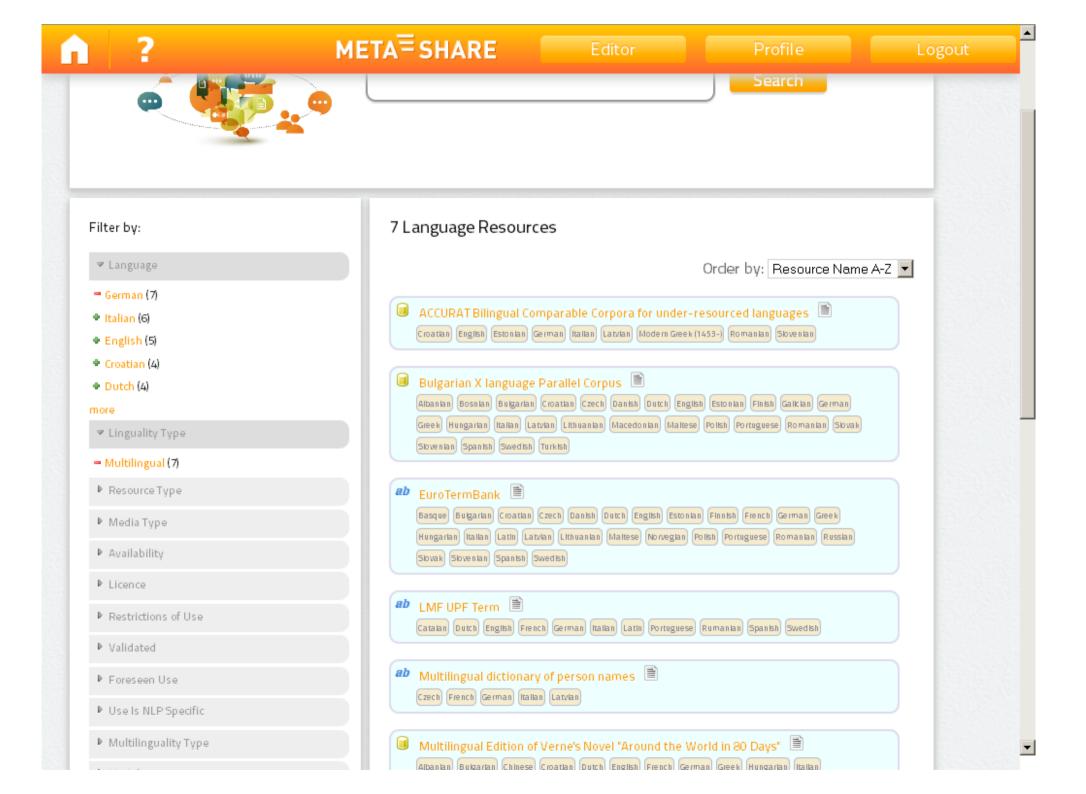






What is it? - About the project

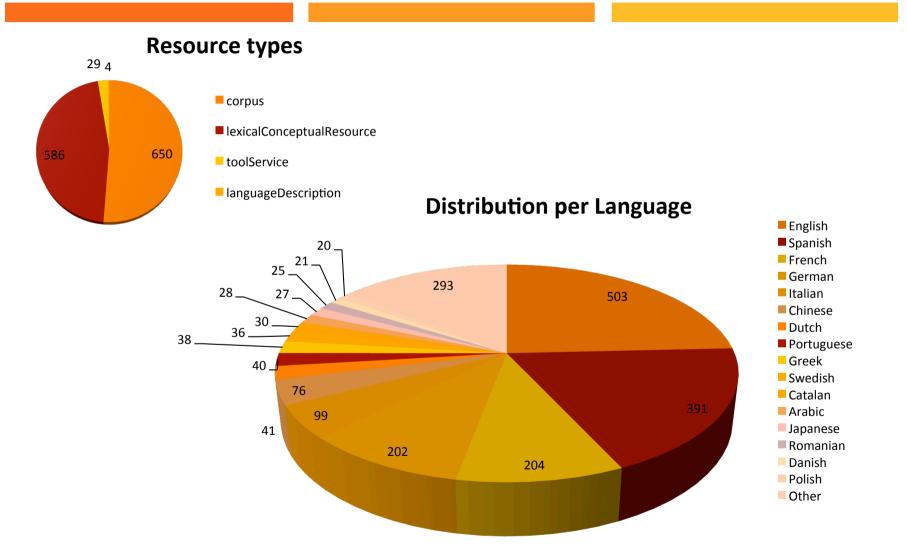
META-NET aims at creating META-SHARE, a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted access rights, free and for a fee META SHARE targets existing but also new and emerging language



?	META	SHARE	Editor	Profile	Logout
Tok Tokeniz	er				
« Back	Download				3 2 √ 1
🗖 Only show manda	tory fields				
	identificatio	ninfo			
Resou	rceName JTok Tokenize				
	scription JTok provides		dentifies paragraphs, sen eviations, numbers, punct	tences and tokens of an input t :uation and clitics.	ext. Non-word
ResourceSt	ortName JTok				
	Url http://hearto	fgold.opendfki.de/brow	ser/trunk/jtok		
Me	aShareld NOT_DEFINE	D_FOR_V2			
	distribution	Info			
A	vailability available-unr	estrictedUse			
	licenceInfo				
	Licence LGPL				
DistributionAcces		e			
Downloa		fgold.opendfki.de/brow	ser/trunk/jtok		
	contactPers	on			
	Surname Steffen				
	venName Jörg				
	Sex male				

Current Contents





Recent Developments



- Several organizations outside META-NET want to set up their own META-SHARE repositories (e.g., JRC Optima Activity).
- Incoming resources from CESAR, METANET4U and META-NORD and also from additional initiatives (collaboration agreements with >40 EU-funded projects).
- ELRA will adopt META-SHARE as the central model for LR sharing and will provide services in its operation.
- META-SHARE Version 3.0 to be released in September 2012.



META-VISION

Language White Paper Series

http://www.meta-net.eu

Language White Papers 1/2



- Key communication instruments to address decision makers and journalists.
- Cover all EU languages (30 volumes).
- Report on the state of a language (general, social, strategic and technological aspects) and the level of support through language technology.
- Inform target group about societal and technological problems and challenges as well as economic opportunities.
- Printed documents published by Springer; will be distributed by META-NET.
- PDF versions available for free.



30 Languages Covered



- Basque
- Bulgarian*
- Catalan
- Czech*
- Danish*
- Dutch*
- English*
- Estonian*
- Finnish*
- French*

- Galician
- □ German*
- Greek*
- Hungarian*
- Icelandic
- Irish*
- Italian*
- Latvian*
- Lithuanian*
- Maltese*

- Norwegian
- Polish*
- Portuguese*
- Romanian*
- Serbian
- Slovak*
- Slovene*
- Spanish*
- Swedish*
- Croatian

* = Official EU language

META⁼ NET

White Paper Series Séria bielych kníh

THE SLOVAK SLOVENSKÝ LANGUAGE IN JAZYK THE DIGITAL V DIGITÁLNOM AGE VEKU

Mária Šimková Radovan Garabík Katarína Gajdošová Michal Laclavík Slavomír Ondrejovič Jozef Juhár Ján Genči Karol Furdík Helena Ivoríková Jozef Ivanecký

🖄 Springer

Ξ

OBSAH CONTENTS

SLOVENSKÝ JAZYK V DIGITÁLNOM VEKU

1	Zhrnutie	1
2	- ··· ··· - ··· ··· ··· ··· ··· ···	3
	2.1 Jazykové hranice spomaľujú európsku informačnú spoločnosť	4
	2.2 Naše jazyky v ohrození	4
	2.3 Jazykové technológie sú kľúčovými technológiami	5
	2.4 Príležitosti pre jazykové technológie	5
	2.5 Výzvy pre jazykové technológie	6
	2.6 Osvojovanie si jazyka	6
3	Slovenčina v európskej informačnej spoločnosti	8
	3.1 Všeobecné fakty	8
	3.2 Špecifiká slovenčiny	11
	3.3 Slovenčina na internete	12
	3.4 Slovenčina ako cudzí jazyk	13
	3.5 Slovenský národný korpus	15
4	Jazykové technológie na podporu slovenčiny	17
	4.1 Architektúra aplikácií	17
	4.2 Základné aplikačné oblasti	19
	4.3 Ďalšie aplikačné oblasti	27
	4.4 Jazykové technológie vo vzdelávaní	29
	4.5 Štátne programy a iniciatívy	29
	4.6 Dostupnosť nástrojov a zdrojov	30
		32
	4.8 Závery	
5	O META-NET-e	37

THE SLOVAK LANGUAGE IN THE DIGITAL AGE

1 Executive Summary

		39

2	Lang	guages at Risk: a Challenge for Language Technology	41			
	2.1	Language Borders Hold back the European Information Society	42			
	2.2	Our Languages at Risk	42			
	2.3	Language Technology is a Key Enabling Technology	42			
	2.4	Opportunities for Language Technology	43			
	2.5	Challenges Facing Language Technology	44			
	2.6	Language Acquisition in Humans and Machines	44			
3	Slov	ak in the European Information Society	46			
	3.1	General Facts	46			
	3.2	Particularities of the Slovak Language	49			
	3.3	Slovak on the Internet	51			
	3.4	Slovak as a Foreign Language	51			
	3.5	Slovak National Corpus	53			
4	Lang	guage Technology Support for Slovak	55			
	4.1	Application Architectures	55			
	4.2	Core Application Areas	57			
	4.3	Other Application Areas	65			
	4.4	Language Technology in Education	66			
	4.5	National Projects and Initiatives	67			
	4.6	Availability of Tools and Resources	67			
	4.7	Cross-language Comparison	70			
	4.8	Conclusions	70			
5	Abo	ut META-NET	74			
A	Zozi	nam literatúry – References	75			
В	B Členovia META-NET-u – META-NET Members 8					

A few Numbers ...



- >160 national experts contributed as authors or co-authors (ca. 5 per language on average)
- >50 additional experts have contributed data and information
- >8.000 copies will be printed and distributed by META-NET



Cross-Lingual Ranking



- Journalists and politicians need simple and clear messages.
- In four application areas, each language is assigned to one of five clusters, ranging from *excellent LT support* to *weak/no support*:
 - **1.** Machine Translation
 - 2. Speech Processing
 - **3.** Text Analysis
 - **4.** Resources
- Results finalised at a meeting in Berlin with representatives of all 30 languages (October 21/22, 2011).



MT (top) & Speech Processing (bottom) META NET



excellent	good	moderate	fragmentary		weak or no support	
	English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian		Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish	
excellent	good	moo	derate		fragmentary	weak or no support
	English	Finnish German	, Dutch, 1, French, n, Italian, se, Spanish	Ca Est Gre Iri Po	sque, Bulgarian, atalan, Danish, onian, Galician, eek, Hungarian, sh, Norwegian, olish, Serbian, ovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian

http://www.meta-net.eu

Text Analysis (top) & Resources (bottom) META NET



excellent	good	moderate	moderate fragmentary		W	zeak or no support
	English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish		La	roatian, Estonian, Icelandic, Irish, atvian, Lithuanian, Maltese, Serbian
excellent	good	moderate		fragmentary		weak/no support
	English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish		Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene		Icelandic, Irish, Latvian, Lithuanian, Maltese

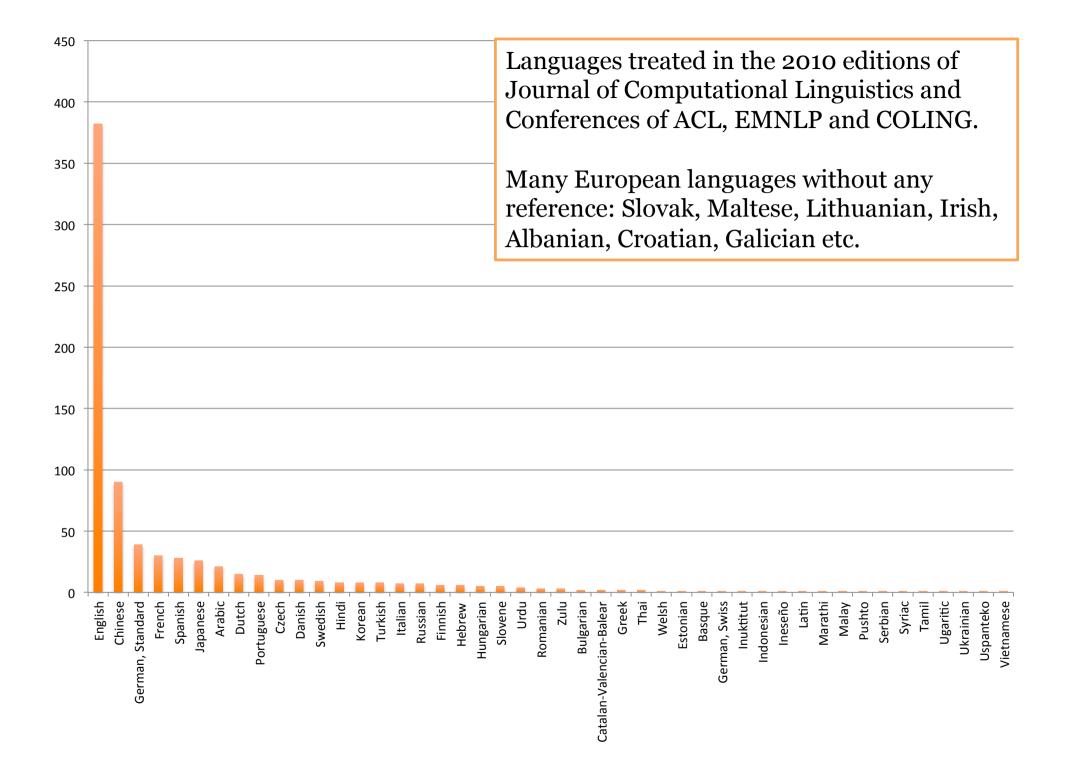
http://www.meta-net.eu

Europe's Languages and LT



English	Dutch French German Italian Spanish	Catalan Czech Finnish Hungarian Polish Portuguese Swedish	Basque Bulgarian Danish Galician Greek Norwegian Romanian Slovak Slovene	Croatian Estonian Icelandic Irish Latvian Lithuanian Maltese Serbian
---------	---	---	--	---

good support through	weak or
Language Technology	no support



Observations



- When it comes to Language Technology support, there are massive differences between Europe's languages and technology areas.
- LT support for English is ahead of any other language but far from being perfect.
- For 16 of 30 languages, LT support is only fragmentary or very weak!

 Final versions available at http://www.meta-net.eu/whitepapers by META-FORUM 2012 (June 20/21, 2012).



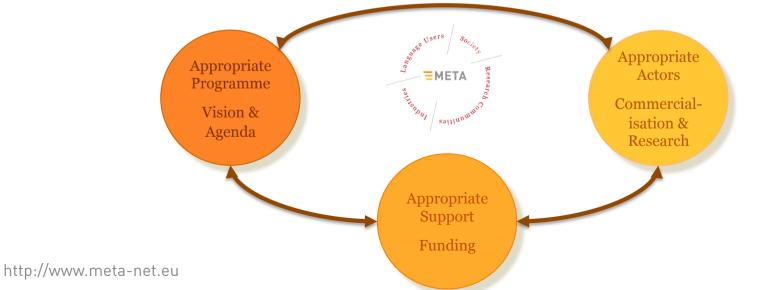
META-VISION

Strategic Research Agenda

Shared Vision and SRA



- Mobilize researchers, users and providers of LT for cooperation and collaboration to initiate and support a large-scale joint action to
 - Building a **community** around Language Technology in Europe (META),
 - Creating a **shared vision**,
 - Preparing a Strategic Research Agenda for Multilingual Europe 2020, to be presented to national/international politicians, funding agencies, decision makers.



From Visions to the SRA



- Three **Vision Groups** brought together researchers, developers, integrators and (corporate or professional) users of LT-based products, services and applications (ca. 25 members each).
- Collected domain-specific visions and prepared individual reports.

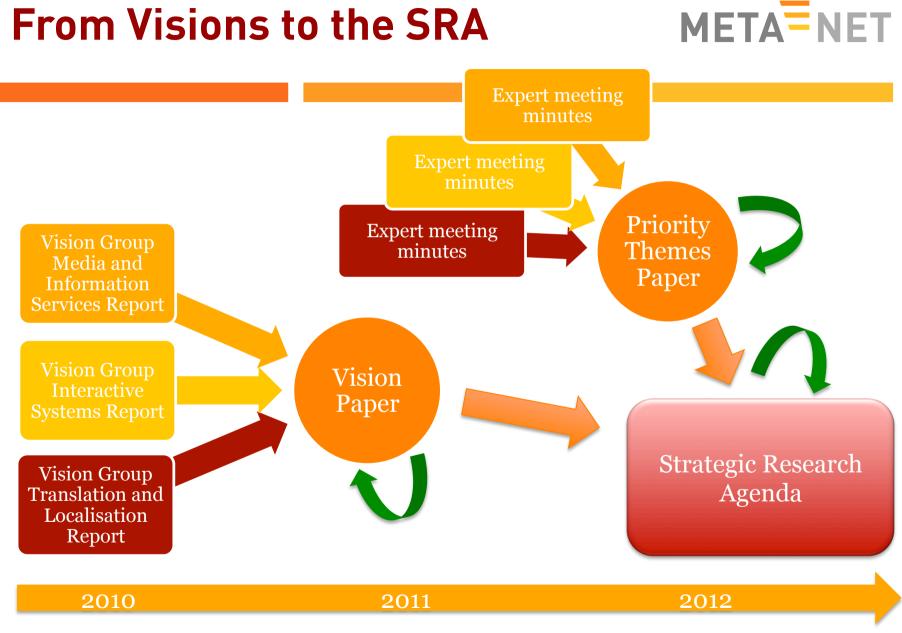
- Vision Group **Translation and Localisation**
- **Vision Group Media and Information Services**
- **Vision Group Interactive Systems**

- July 23, 2010
- September 28, 2010
- April 7/8, 2011
- Berlin, Germany
- Brussels, Belgium
- Prague, Czech Republic
- September 10, 2010
 - Paris, France
 - October 15, 2010 Barcelona, Spain
 - April 1, 2011
- Vienna, Austria
- September 10, 2010 Paris, France October 5, 2010
- March 28, 2011
- Prague, Czech Republic Utrecht, The Netherlands

Steps taken towards the SRA

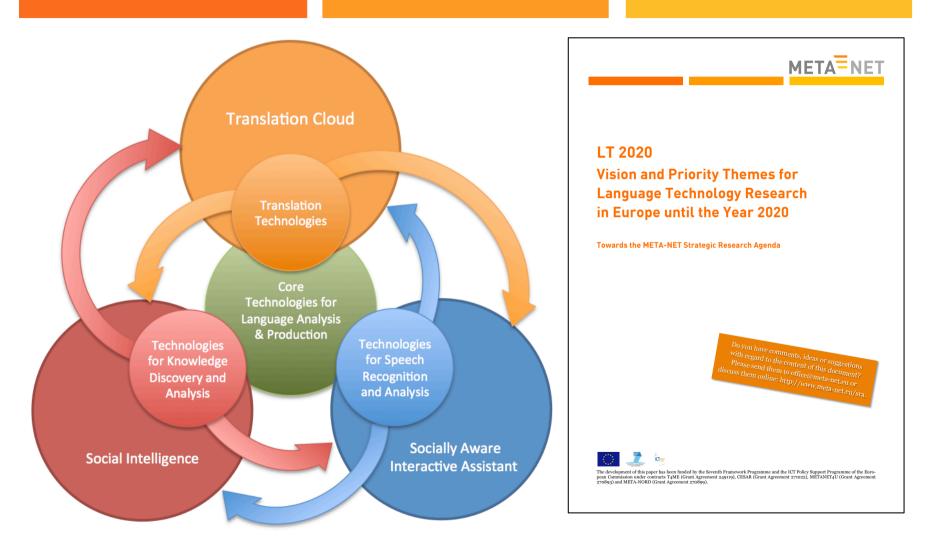


- Many suggestions by: Vision Group members; META-NET; META.
- Additional input in meetings, workshops, discussions etc.
- Language White Papers: research gaps; funding situation etc.
- We screened the Strategic Research Agendas of other initiatives.
- We discussed procedures, input and structure of the SRA in three meetings of the META Technology Council.
 - Brussels, Belgium, November 16, 2010
 - Venice, Italy, May 25, 2011
 - Berlin, Germany, September 30, 2011
 - *Upcoming:* Brussels, Belgium, June 19, 2012



Priority Research Themes

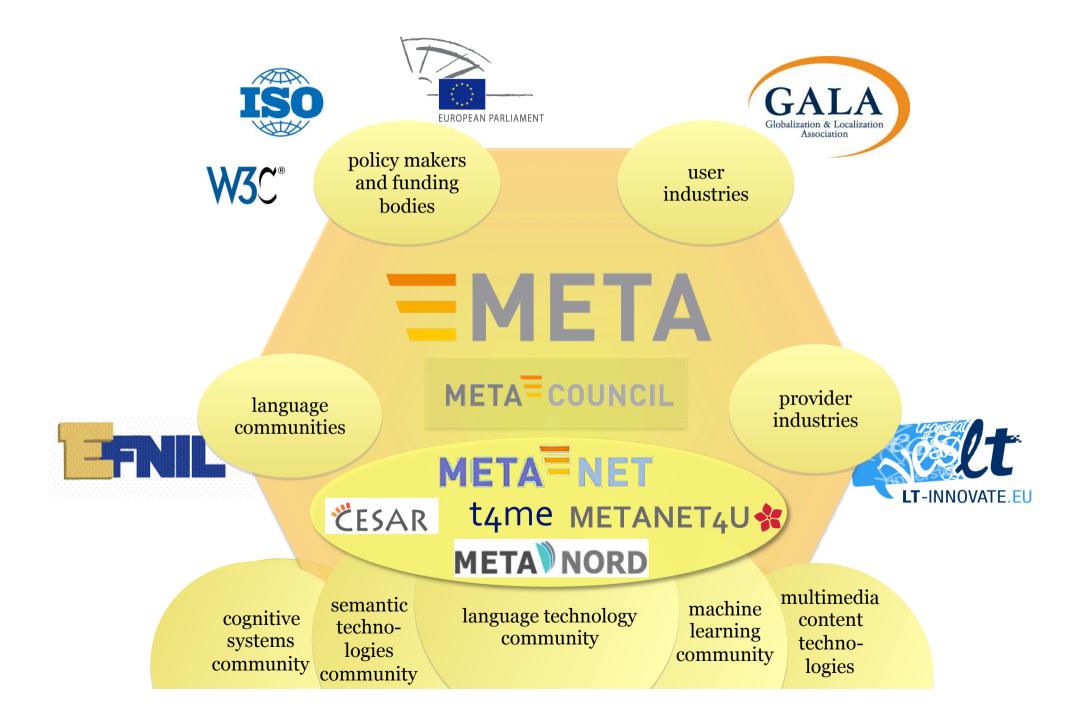






- **META-NET** is a network of excellence.
- META is an open and growing strategic technology alliance:
 Multilingual Europe Technology Alliance.
 - 600+ members, including W3C, Google, Microsoft, GALA, research centres, LT companies etc.
 - Main goal: to support our Strategic Research Agenda.
 - Join us! http://www.meta-net.eu/join





Next Steps



- **•** Roll out META-SHARE for the whole LT community in Sep. 2012.
- Disseminate the Language White Papers widely.
- Meet with national research planners, funders, policy makers and inform them about the Strategic Research Agenda.
- Mobilize user industries and administrations.
- Upcoming opportunities can provide sufficient resources to make our visions for Europe's citizens and economy, as described in the SRA, a reality (Horizon 2020 and Connecting Europe Facility, CEF).



Thank you very much!

See you in Brussels on June 20/21 for

META FORUM 2012

http://www.meta-forum.eu

office@meta-net.eu

http://www.meta-net.eu http://www.facebook.com/META.Alliance