Highlights, Holes *in* and Hopes *for* the "Multilingual Web"

Jan Nelson, Microsoft; Christian Lieske, SAP AG
Felix Sasaki, DFKI; Richard Ishida, W3C

Presenters

Jan Nelson, Microsoft Corporation

- Senior Program Manager in the Windows Division
- Focuses on world-readiness for Windows and partner products
- Active in cross-company internationalization engineering leadership
- Member of the W3C Multilingual Web-LT Working Group
- Holds several patents related to language processing and translation engineering
- One current project: Multilingual App Toolkit (MAT) for developers of Metro style apps running on Windows 8

Christian Lieske, SAP AG

- Knowledge Architect at SAP Language Services
- Involved in content engineering, text processing and process automation (including evaluation, prototyping and piloting)
- Main fields of interest: Internationalization, translation approaches and natural language processing
- Holds several patents related to content engineering
- Contributor to standardization at World Wide Web Consortium (W3C), OASIS

Outline

Supporting the Multilingual Web

- **Running a Thematic Network**

Working *for* and *in* the Multilingual Web

- **Laying foundations and building on them**

Discussion

What is today‘s Web?

Fundamental to communication in all walks of life

+

Backbone for economy and government

+

Set of base technologies

→

Something that needs to support any culture and language
=
Something that needs to be **multilingual**

"Creation, preservation and processing of, and access to [..] content in digital form should [..] **ensure that all cultures can express themselves and have access to Internet in all languages,** including indigenous and minority languages." (UNESCO, Code of Ethics for the Information Society (Draft))

How ensure support for any culture and language?

Provide insights into standards, best practices and gaps

Organize exchange and networking between stakeholders

Run a Cooperation and Support Action (CSA)

- **Funded by the European Commission (EC)**
- **Coordinated by the World Wide Web consortium (W3C)**
- **Driven by partners from industry, academia, organizations and others**

Gather for events in beautiful places such as Madrid, Pisa, Limerick, and Luxembourg ☺

Six major stakeholders and areas

Developers

- Organizations such as the W3C, the Internet Engineering Task Force (IETF), the Unicode Consortium, …
- Major implementers of user agents (e.g. browsers)
- …

Creators

- You ☺
- News providers
- Web developers
- Large and small companies
- …

Localizers

- Translation agencies
- Localization researchers
- Localization departments of large companies

Machines

- Natural Language Proccessing companies and researchers
- Enablers for new markets via new services, e.g. sentiment analysis

Users

- You ☺
- People with special needs
- Smaller and larger language communities

Policy makers

- Governments fighting against language barriers & for new market opportunities
- Enthusiasts forming innovative research topics & strategies

The stakeholders and areas are connected.
Some topics show up in more than one area.

Caveat

The remainder of the talk will only scratch the surface of the topics that were covered in the different areas

In order to learn more, please visit www.multilingualweb.eu

The „Events&Reports“ section of the site provides access to

- **Event summaries**
- **Transcripts taken by scribes during the meeting**
- **Video recordings (where available)**
- **Social media links (e.g. to blog postings)**

It is strongly recommended to watch the videos, where available, since these are short but carry much more detail.

Developers – Laying or enhancing foundations

Word clouds generated by wordle.net

- Support more scripts
- Encode information on language, script, region, locale,
- Support vertical direction
- Allow ruby annotations
- Enhance right-to-left and bidirectional
- Enable rendering of complex scripts
- Handle dates, calendars, time zones

Developers – Criticism raised to them

- Standards too abstract, too heavy, and too numerous
- Outreach and education insufficient
- Collaboration between constituencies limited
- Partipication options restricted
- Supply chain processes not addressed
- User preferences insufficiently supported
- Multilingual Web Sites not harmonized
- Long tail cultures and languages without priority
- Testability insufficient
- Reference implementations lacking
- Implementations not compliant or interoperable

Developers – Suggestions provided to them

- Create Internet Assigned Numbers Authority (IANA) Registry for Phonetic Alphabets
- Allow binding of events to XML
- Add part-of-speech tags to Pronunciation Lexicon Specification (PLS)
- Work on indicators for provenance (e.g. „has been machine translated“)
- Work on more powerful, standardized user models
- Address base taxonomy of content states
- Realize that W3C „range“ is insufficient for alignment on word-level

Developers – Their response

The Web needs your help. This is your Web. The Web is about people, not technology. We need you to make the Web worldwide. Don't rely on us to do the work for you!

- Follow the discussions on the i18n mailing lists (eg. www-international@w3.org), and track other technologies for internationally relevant topics. Follow our RSS feeds and twitter channels (@webi18n and @multilingweb)
- Read and review specifications (http://www.w3.org/TR/tr-technology-drafts) and send comments to the i18n list or direct to the Working Group.
- Discuss local requirements for the Multilingual Web, and if you identify missing features, find ways to coordinate proposals.
- Use features needed for non-Latin script support and push implementers to include more in browsers and authoring tools.
- Review or contribute to development/dissemination of outreach materials, to help others understand how to implement and use international features of the Web.
- Take on board that internationalization is something done by developers and designers – not localizers. Find out how to do it. (http://www.org/International/)
- Use Unicode and UTF-8 wherever you can.
- Consider how your content will appear on the Mobile Web.
- Use the I18n Checker (http://qa-dev.w3.org/i18n-checker/) and send ideas for improvements.

Creators – Putting content on the web

Reality

- 23 mio unique visitors per week (23 languages)
- 1 billion visits per year (40 languages)
- Demand for personalized, and thus often locally relevant content
- Real-time information delivery (simulatenously across languages)
- Several scripts for single language (e.g. Uzbek)
- 70% of mobile devices cannot display Hindi correctly
- Explosion of volume, decrease in granularity
- Integration of proprietary systems

Pains

- Lacking implementations, interoperability and compliance
- Missing font support
- Missing support for input/keyboards
- Inter-language and cross-country links
- Limited connection speed/bandwith

Creators – Suggestions to/requirements for them

Serve more modalities than text

Assure via mobile approaches that everybody can be online

Create proper multilingual links

Make local content available world-wide

Take requirements from the content domain (e.g. health care) into account

Help to bring quality translation with volunteers into Wikipedia

Creators – Current directions

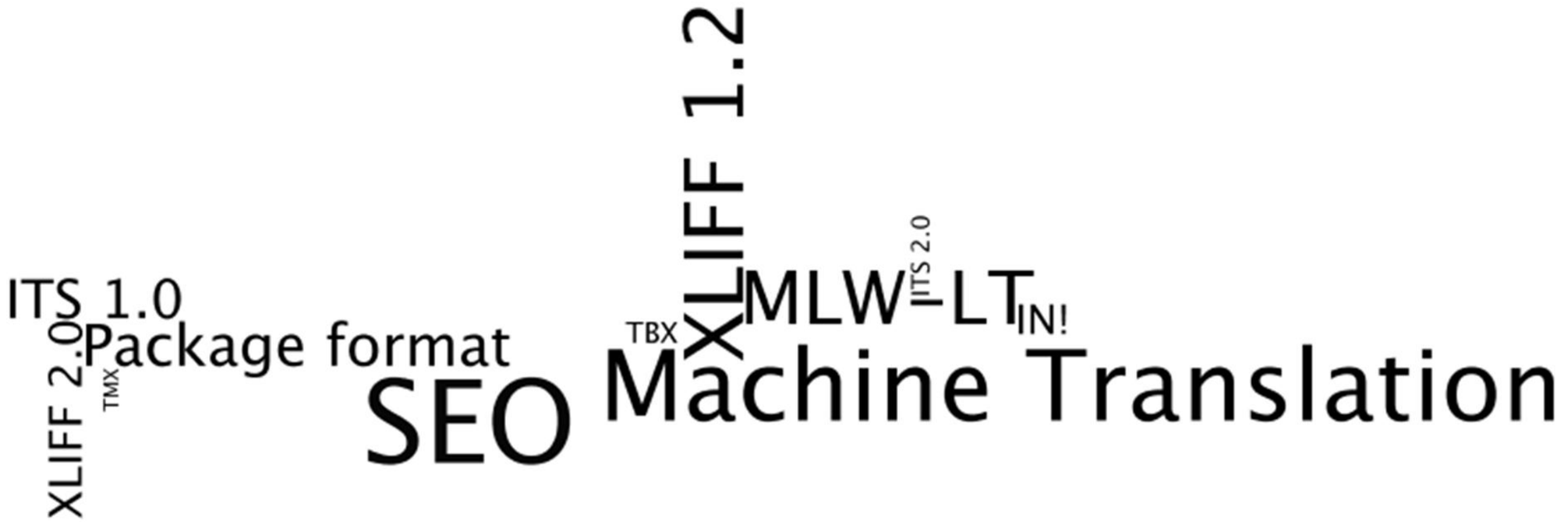
Workarounds

- Deliver text as image
- Set up country-/language-specific content-related workflows

New approaches

- Use office applications to create best practice HTML/XML
- Apply „English as just another language“ paradigm
- Crowd-source content creation and/or translation
- Crowd-source translation of solution itself
- Provide tooling for translation, and reviewing
- Offer on-the-fly machine translation
- Consider best practices: docs.webplatform.org/wiki/Main_Page

Localizers – Helping to speak the language of locals



- **Immediacy**
- **User-generated content (including Social Media contributions)**
- **Changing service requests/business models**
- **Auto-completion and sub-segment matching**
- **Long tail cultures and languages**
- **Tools and incentives provided by „creators“**

Localizers – Pains and opportunities

Pains

- Standards for content creation don’t take localization needs into account
- Localization standards abound
- Localization misses interoperable implementations
- Automation in localization is missing quality assessment/provenance information

Opportunities

- Qualit-aware crowd-sourcing can bring the Web to all language communities
- Small, implementation-driven steps in localization standards/interoperability
- HTML5!

Localizers – Suggestions to them

- Think twice before starting a localization standard, or relying on one
- Match localization workflows, content creators and project needs
- Be brave: use language technology
- Be brave: localize in the browser, and embrace the cloud
- Help languages with limited resources, and consider all modalities
- Use and generate open data in localization processes

Localizers – Tools from „creators“

MS Collaborative Translation Framework	Combine automatic machine translation with human translation
	http://blogs.msdn.com/b/translation/archive/2010/03/15/collaborative-translations-announcing-the-next-version-of-microsoft-translator-technology-v2-apis-and-widget.aspx
MS Multilingual App Toolkit (MAT)	Localize Windows Store app
	http://msdn.microsoft.com/en-us/windows/apps/br229516
WikiBasha Beta	Create multilingual content for Wikipedia
	http://www.wikibhasha.org
Pontoon	Localize web site live
	https://wiki.mozilla.org/L10n:Pontoon
BE-COLA	Localize Web Content in Micro Crowdsourcing Architecture and build Translation Memories
	http://videolectures.net/w3cworkshop2011_wasala_web/
LetsMT!	Build and run your own custom machine translation systems
	https://www.letsmt.eu
Narayam	Add and manage different language input methods (for WikiMedia)
	http://www.mediawiki.org/wiki/Extension:Narayam
Translatewiki.net	Localize
	http://translatewiki.net/wiki/Main_Page

Machines – Enabling more and more efficient language-related offerings

- Summarization
- Data-driven/statistical Machine Translation
- Text mining
- Text classification
- Enrichment
- Harvesting and cleansing
- Crosslingual access and processing
- Improved content management
- Improved text-based analytical solutions

Machines – Focus and directions

- Interoperability between technologies
- Reuse and integration of isolated language resources
- Approaches for languages with limited resources
- Identification and annotation based on Semantic Web principles
- Web-based resource creation and coupling of processing
- Abstraction from language/lexical level to conceptual level
- Mathematical/statistical approaches like correlated vector spaces
- Ontology-based processes
- Engine factories

Users – Demanding more than English and contributing more

General expectations

- Real-time
- Personalized (especially translated and locally relevant)
- Transparent, complete, consistent (content, links, site maps, indices)
- Accessible

Example *Facebook*

- 75% growth of user base outside of US in 2010
- 500% increase overnight in use of Arabic User Interface via easy language selection mechanism
- 500000 voluntary translators; French translated within 24 hours

Other examples

- Kiswahili Wikipedia content creation, and high quality health information facilitated by Google

Policy Makers – Mandating and promoting

Technology at service of society

- Multilingual mandates, participatory democracy
- Interactive systems for local needs
- Open multilingual assets (e.g. legal and administrative terminology)
- Harmonize support (23 out of 30 European language suffer from limited Machine Translation)
- Education, promotion, coordination, guidelines, business cases related to multiliguality on the web

Governments can influence by mandating standards (e.g. use of Unicode for any type of persistency)

Conclusions

The web needs to be multilingual

Support for more multilinguality on the web is on its way

More support is needed – Your‘s!

Thank You!

Contact information:

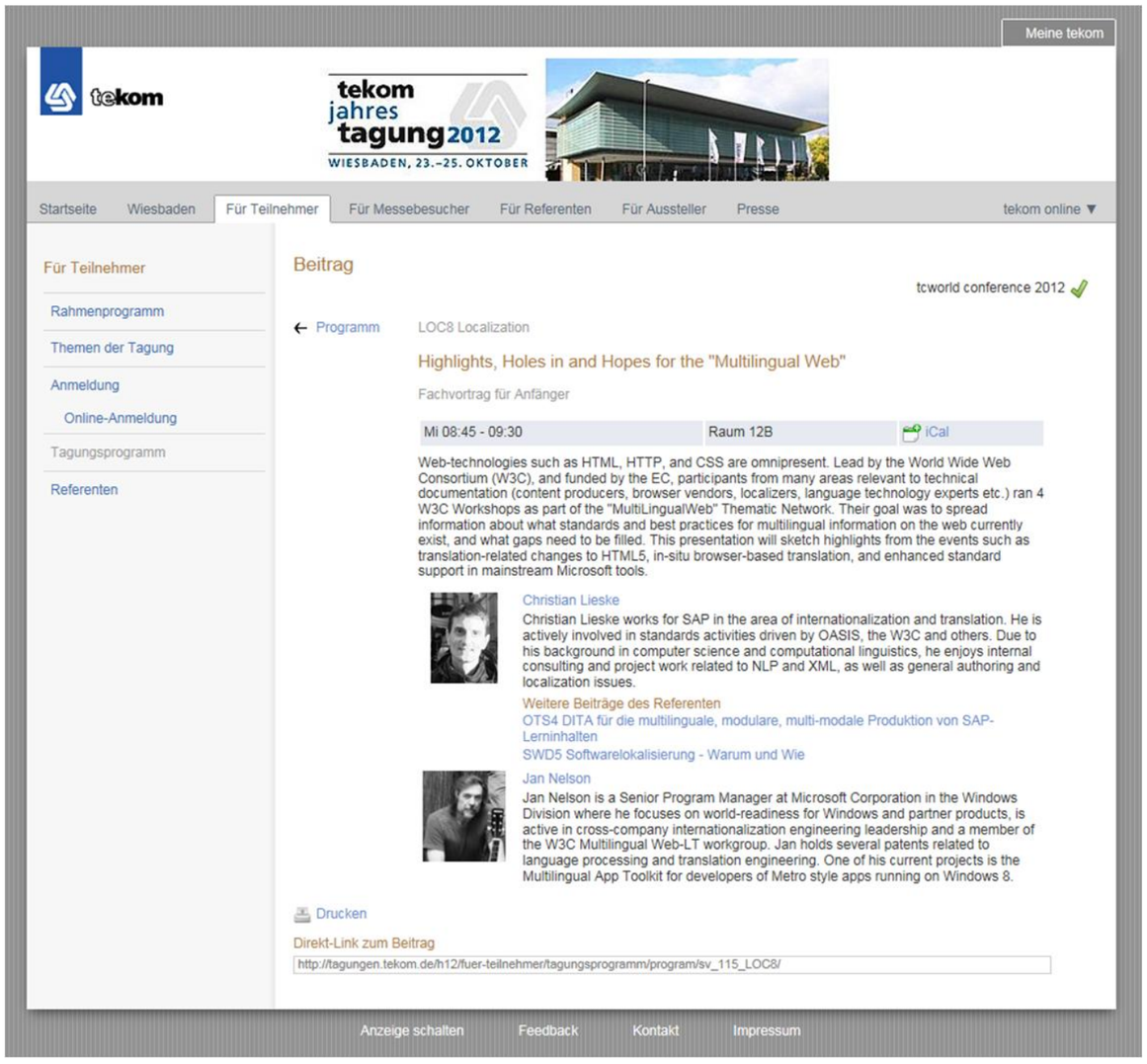
Christian Lieske
christian.lieske@sap.com
www.sap.com

Felix Sasaki
felix.sasaki@dfki.de
www.dfki.de

Jan Nelson
Jan.Nelson@microsoft.com
www.microsoft.com

Richard Ishida
ishida@w3.org
www.w3.org

Appendix



http://tagungen.tekom.de/h12/fuer-teilnehmer/tagungsprogramm/program/sv_115_LOC8/

Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, ohne die ausdrückliche schriftliche Genehmigung durch SAP AG nicht gestattet. In dieser Publikation enthaltene Informationen können ohne vorherige Ankündigung geändert werden.

Die von SAP AG oder deren Vertriebsfirmen angebotenen Softwareprodukte können Softwarekomponenten auch anderer Softwarehersteller enthalten.

Microsoft, Windows, Excel, Outlook, PowerPoint, Silverlight und Visual Studio sind eingetragene Marken der Microsoft Corporation.

IBM, DB2, DB2 Universal Database, System i, System i5, System p, System p5, System x, System z, System z10, z10, z/VM, z/OS, OS/390, zEnterprise, PowerVM, Power Architecture, Power Systems, POWER7, POWER6+, POWER6, POWER, PowerHA, pureScale, PowerPC, BladeCenter, System Storage, Storwize, XIV, GPFS, HACMP, RETAIN, DB2 Connect, RACF, Redbooks, OS/2, AIX, Intelligent Miner, WebSphere, Tivoli, Informix und Smarter Planet sind Marken oder eingetragene Marken der IBM Corporation.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und anderen Ländern.

Adobe, das Adobe-Logo, Acrobat, PostScript und Reader sind Marken oder eingetragene Marken von Adobe Systems Incorporated in den USA und/oder anderen Ländern.

Oracle und Java sind eingetragene Marken von Oracle und/oder ihrer Tochtergesellschaften.

UNIX, X/Open, OSF/1 und Motif sind eingetragene Marken der Open Group.

Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame und MultiWin sind Marken oder eingetragene Marken von Citrix Systems, Inc.

HTML, XML, XHTML und W3C sind Marken oder eingetragene Marken des W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.

Apple, App Store, iBooks, iPad, iPhone, iPhoto, iPod, iTunes, Multi-Touch, Objective-C, Retina, Safari, Siri und Xcode sind Marken oder eingetragene Marken der Apple Inc.

IOS ist eine eingetragene Marke von Cisco Systems Inc.

RIM, BlackBerry, BBM, BlackBerry Curve, BlackBerry Bold, BlackBerry Pearl, BlackBerry Torch, BlackBerry Storm, BlackBerry Storm2, BlackBerry PlayBook und BlackBerry App World sind Marken oder eingetragene Marken von Research in Motion Limited.

Google App Engine, Google Apps, Google Checkout, Google Data API, Google Maps, Google Mobile Ads, Google Mobile Updater, Google Mobile, Google Store, Google Sync, Google Updater, Google Voice, Google Mail, Gmail, YouTube, Dalvik und Android sind Marken oder eingetragene Marken von Google Inc.

INTERMEC ist eine eingetragene Marke der Intermec Technologies Corporation.

Wi-Fi ist eine eingetragene Marke der Wi-Fi Alliance.

Bluetooth ist eine eingetragene Marke von Bluetooth SIG Inc.

Motorola ist eine eingetragene Marke von Motorola Trademark Holdings, LLC.

Computop ist eine eingetragene Marke der Computop Wirtschaftsinformatik GmbH.

SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, SAP HANA und weitere im Text erwähnte SAP-Produkte und -Dienstleistungen sowie die entsprechenden Logos sind Marken oder eingetragene Marken der SAP AG in Deutschland und anderen Ländern.

Business Objects und das Business-Objects-Logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius und andere im Text erwähnte Business-Objects-Produkte und Dienstleistungen sowie die entsprechenden Logos sind Marken oder eingetragene Marken der Business Objects Software Ltd. Business Objects ist ein Unternehmen der SAP AG.

Sybase und Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere und weitere im Text erwähnte Sybase-Produkte und -Dienstleistungen sowie die entsprechenden Logos sind Marken oder eingetragene Marken der Sybase Inc. Sybase ist ein Unternehmen der SAP AG.

Crossgate, m@gic EDDY, B2B 360 ° , B2B 360 ° Services sind eingetragene Marken der Crossgate AG in Deutschland und anderen Ländern. Crossgate ist ein Unternehmen der SAP AG.

Alle anderen Namen von Produkten und Dienstleistungen sind Marken der jeweiligen Firmen. Die Angaben im Text sind unverbindlich und dienen lediglich zu Informationszwecken. Produkte können länderspezifische Unterschiede aufweisen.

Die in dieser Publikation enthaltene Information ist Eigentum der SAP. Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, nur mit ausdrücklicher schriftlicher Genehmigung durch SAP AG gestattet.

Disclaimer

All product and service names mentioned and associated logos displayed are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

This document may contain only intended strategies, developments, and is not intended to be binding upon the authors or their employers to any particular course of business, product strategy, and/or development. The authors or their employers assume no responsibility for errors or omissions in this document. The authors or their employers do not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.

The authors or their employers shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.

The authors have no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages.

Partners in the *Multilingual Web* Thematic Network

Industry

- Bioloom Group, Germany
- Facebook, Ireland
- Language Technology Centre, UK
- Lionbridge, Belgium
- Microsoft Ireland
- Opera Software, Norway
- SAP, Germany
- Transware Ltd (WeLocalize), Ireland
- XML-INTL, UK

Academia

- Aalto-Korkeakoulusaatio, Finland
- Consiglio Nazionale delle Ricerche, Italy
- University of Applied Sciences (UAS) Potsdam, Germany
- Institut Josef Stefan, Slovenia
- Institutul de Cercetari Pentru Intelegentia Articificiala (RACAI), Romania
- University of Oviedo (ILTO), Spain
- Universidad Politécnica de Madrid (UPM), Spain
- University of Economics, Prague, Czech Republic

Standardization Organizations/Other

- European Commission, Directorate-General for Translation, Luxembourg
- Language Resource Centre (LRC), Ireland
- LISA, Switzerland
- Translation Automation User Society (TAUS), Netherlands
- W3C/ERCIM, France (coordination)