The Bricks to Build Tomorrow's Translation Technologies and Processes

Christian Lieske (SAP AG), Felix Sasaki (DFKI), Yves Savourel (ENLASO)

W3C Workshop: Content on the Multlingual Web, 4-5 April 2011, Pisa

Agenda

1. Why talk about tomorrow’s Translation Technologies and Processes?
2. What are the most essential Ingredients for building the Tomorrow?
3. Outlook

Introductory Remarks

„Bricks“ is misleading since it refers to static entities – the *What?*

At the current point in time, focus should be on dynamic entities (namely mindsets, and approaches) – the *How*?

In addition, to „bricks“, the overall architecture needs to be considered.

Presenter

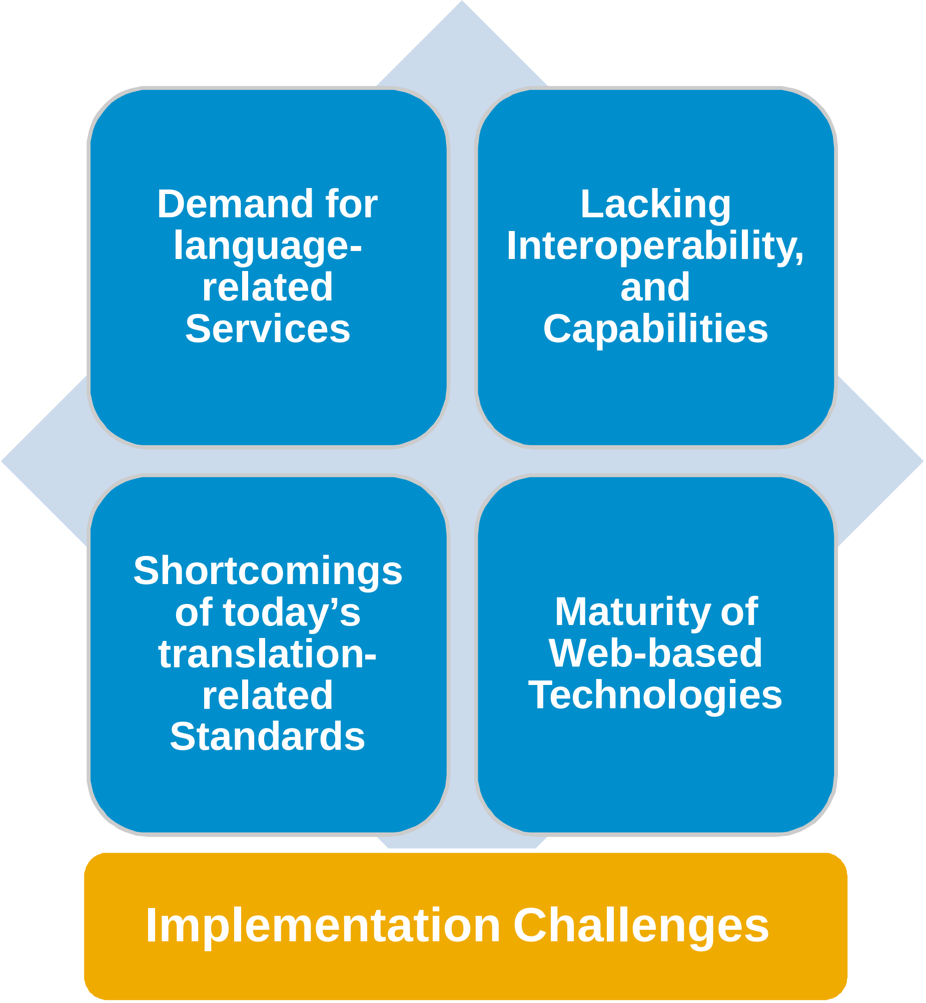
Christian Lieske

SAP Language Services
Globalization Services
SAP AG

- Knowledge Architect
- Content engineering and process automation (including evaluation, prototyping and piloting)
- Main field of interest: Internationalization, translation approaches and natural language processing
- Contributor to standardization at World Wide Web Consortium (W3C) OASIS and elsewhere
- Degree in Computer Science with focus on Natural Language Processing and Artificial Intelligence

This presentation is **purely personal** — our employers have no responsibility for any information contained here

Why talk about tomorrow’s Translation Technologies and Processes?



Demand for language-related Services

Lacking Interoperability, and Capabilities

Shortcomings of today’s translation-related Standards

Maturity of Web-based Technologies

Implementation Challenges

Why? – Demand & Lacking Interoperability

1. **There is an ever increasing demand for automated, interoperable translation-/language-related services.**

- Studies from the EC (see "The size of the language industry in Europe" (Adriane Rinsche et al., http://ec.europa.eu/dgs/translation/publications/studies/index_en.htm)
- Statements from Translators without Borders/Rosetta Foundation

2. **Today's automation lacks interoperability, and capabilities.**

- XLIFF implementations
- No official JSON representations for standards
- Missing support for "elementsWithinText" or "translate" in Machine Translation interfaces like bing or google translate

Why? – Shortcomings of Standards & Use Web Technologies

3. **Models are not harmonized and standardized, and thus require substantial efforts to be utilized**

- *seg/trans-unit* in TMX and XLIFF
- Inline markup in TMX and XLIFF
- Missing markup in TBX definitions

3. **Little work has been done on Web technologies (e.g. communication protocols) in translation-related technologies**

- Utilitization of standardized RESTful services
- JavaScript
- Use of OData or GData for queries or updates

Compare to similar movements in other areas like XQuery in the browser (e.g. XML Prague 2011 http://www.xmlprague.cz/2011/index.html)

Why? – Implementation Challenges

5. **Today's translation-related standards are complex and hard to implement**

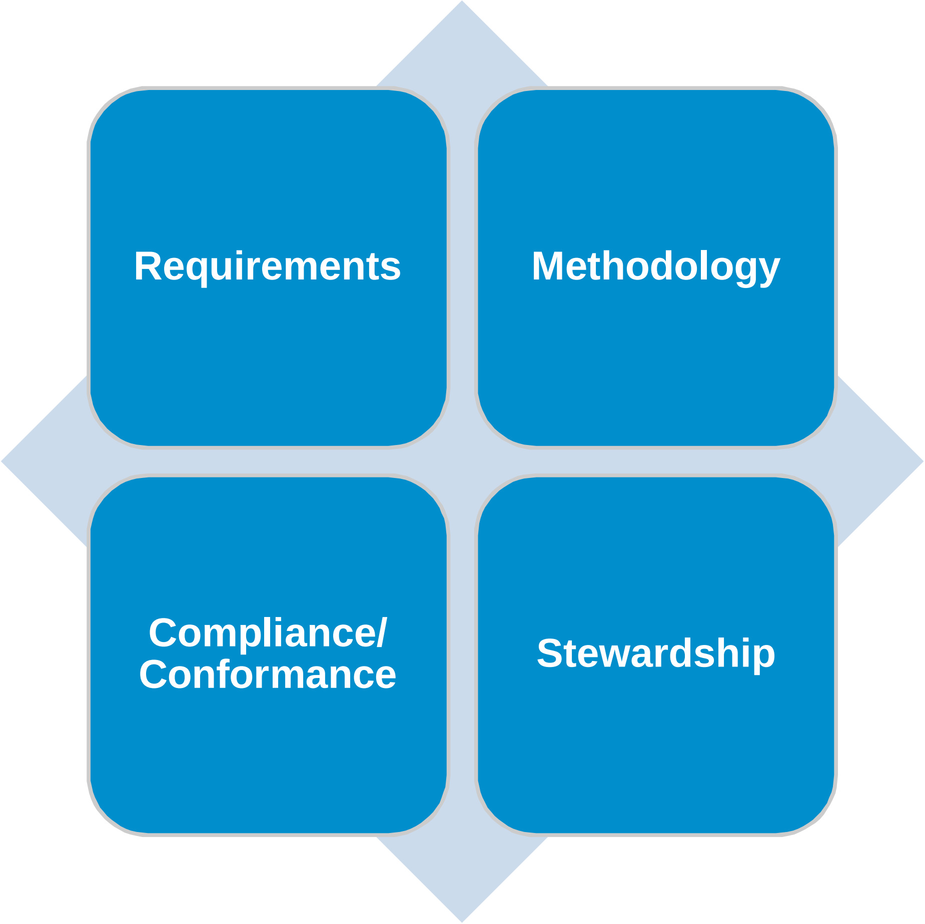
- Insights from First XLIFF Symposium
- Depending on XPath is limitative because it is not implemented everywhere
- Forcing SRX to use ICU regex constructs is bad because it cannot currently be done in Java

2. - 5. result in efficiencies during design time and run time.

You need costly experts to set up processes, and have to do a lot of back and forth conversions.

Example: Couple a database with C++ runtime messages with an online Machine Translation System

What are the most essential Ingredients for building the Tomorrow?



What? – Requirements

1. **Identify processing areas related language processing - and keep them apart**

   Extraction of text units, segmentation, …

2. **Determine the entities that are needed in each area**

   “extraction of text units”: markers to distinguish text from non-text, mechanism to remerge text units with non-text, …

3. **Chart technology options and needs**

   Are RDF/RDFa, OWL – main ingredients of the Semantic Web – viable representation approaches?

4. **Realize opportunities to reuse, and worship standards**

- Use BCP47 for language identifiers (de-DE-u-attr-co-phonebk - "German in phonebook collation order“)
- Tendency for convergence (different technology stacks for Semantic Technologies are more and more being aligned; Semantic Web (RDF or the RDFa serialization), microformats, ...)
- OData/GData as powerful combination based on Atom, AtomPub, HTTP, XML and JSON

In order to maximize synergies and to avoid risk do all of this as transparent as possible.

What? – Methodology

1. **Distinguish between models and implementations/serializations …**

   RDF models/formats (XML, turtle, …)

2. **Distinguish between entities without context and entities with business/processing context**

   Language identifier = without context; source language identifier = with context

3. **Set up rules to transform data models into syntaxes**

   Ensure that the XSD representation for language related concepts always uses *xml:lang*

4. **Set up flexible registries (or even more powerful collaboration tools e.g. to allow composition of new formats from building blocks)**

   Common locale data registry, IANA

Provide migration paths/mapping mechanisms for legacy data

Map from your own approach to *xml:lang* language identification (see W3C ITS)

The Core Components Technical Specification (CCTS) developed within UN/CEFACT, UBL and ebXML exemplify some of the above.

http://www.sdn.sap.com/irj/sdn/index?rid=/webcontent/uuid/27755904-0b01-0010-25b6-bd2629bfa83e

http://www.sdn.sap.com/irj/sdn/go/portal/prtroot/com.sap.km.cm.docs/media/uuid/003216b0-0b6d-2a10-db9b-aa9037feae7e

What? – Compliance

1. **Thou shall have compliance statements**

   Difficult situation with XLIFF (where XLIFF 1.2 does not have compliance clauses)

2. **Thou shall provide test cases (aside: this is far more than test material)**

   W3C ITS, …

3. **Thou shall publish results from test runs if you claim compliance/conformance**

   W3C ITS, Web browser tests

4. **You may mandate proofs of interoperability (possibly even in the disguise of public events)**

   OASIS rules for liasons/ISO fast track; HL7 Connectathon

5. **You may benefit from singleton implementations**

   If all use the same library for reading/writing ...

What? – Stewardship

1. **Realize that resources are needed, need to be connected and coordinated**

   The EC has a track record related to this (see the Multilingual Web Thematic Network)

2. **Make donations/contributions easy**

3. **Discourage fragmentation and unclear roles**

4. **Think out of the box**

   Do not just buddy with colleagues from translation, but also with people who are into Web technologies, language technologies, users, content (tool) providers

3. **Model “same person works in several roles” (W3C, Unicode, OASIS, IETF, ...) works well in certain cases**

4. **Know of pragmatic realities**

   See how e.g. "Moses for Localization" google group (http://groups.google.com/group/m4loc/) establishes de-facto standards

5. **Preserve heritage**

   Unsure what will happen to the formats developed within the Localization Industry Standards Association (LISA)

Thank You!

Contact information:

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

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

Yves Savourel
ysavourel@translate.com
www.translate.com

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.