XLIFF 2.0

Dr. David Filip

Multilingual Web Workshop
CNR PISA, April 4, 2011

Agenda

1. What are the areas of Language Technology (LT) metadata standardization?
2. Which are the natural homes for LT standardization?
3. Why XLIFF, and why 2.0?
4. What are the issues of 1.2? Tweaks..
5. XLIFF 2.0 SWOT Analysis
6. Challenges for 2.0

1. Q&A at the end of the L10n block

David will argue that content metadata must survive language transformations to be of use in multilingual web. In order to achieve that goal, content creation and content langauge transformation related meta-data must be congruent, i.e. designed upfront with the transformation processes in mind. To make the point for XLIFF as the principal vehicle for critical metadata throughout multilingual transformations, it will be necessary to give a high level overview of XLIFF structure and functions, both in the current version and the next generation standard that is currently a major and exciting work in progress in the OASIS XLIFF TC.

David will argue that content **metadata must survive language transformations** to be of use in multilingual web. In order to achieve that goal, **content** creation and content langauge transformation related **meta-data must be** congruent, i.e. **designed upfront with the transformation processes in mind**. To make the point for **XLIFF as the principal vehicle for critical metadata throughout multilingual transformations**, it will be necessary to give a high level overview of XLIFF structure and functions, both in the current version and the **next generation standard** that is currently a **major and exciting work in progress in the OASIS XLIFF TC**.

- **Metadata must survive language transformations**

- **Content meta-data must be designed upfront with the transformation processes in mind**

- **XLIFF is the principal vehicle for critical metadata throughout multilingual transformations**

- **The next generation XLIFF standard is a major and exciting work in progress in the OASIS XLIFF TC**

The factor of **preserving metadata throughout various types of internationalization, localization and translation transformations (manual, automated, assisted etc.; translation, editing, stylistic review, subject matter review, tagging, gisting etc.) will become critical with multiple source languages becoming standard rather than exception in large multilingual content repositories** (current examples: Wikipedia, knowledge bases and community generated support content).

Preserving Metadata

- **Transformation areas**
  - **GILT (G11n, I18n, L10n, T9n)**
- **Transformation modi:**
  - **Manual, Automated, Assisted**
- **Transformation types:**
  - **machine translation, human translation, (post)editing, stylistic review, subject matter review, tagging, transcribing, subtitling, gisting etc.**
- **Growing number of source languages**
- **Large multilingual content repositories**

It is critical to secure semantics match between content creation and transformation processes standards, to marry content creation, localization and publishing standards.

What Meatadata?

data and metadata structures for context preview generation

- reference implementation of standardized xslt preview artifacts that will be designed to facilitate relevant round-trips throughout all human assisted roundtrips within the content life cycle. The business case is immense, ask Dag.
- Skeleton provisions in XLIFF
- XLIFF crucial for preview generation or preview information transfer in a number of tools (WorldServer DWB, Multicorpora XLIFF editor, Alchemy Publisher etc.)

What Meatadata?

metadata for legally conscious sharing, such as ownership, licensing etc.

- Past content was not created for sharing. However, because of exponential context explosion future data is incomparably more important than the past data.
- Future data must be created upfront with sharing in mind. Legal, privacy and Intellectual Property Rights (IPR) related metadata are one of key prerequisites of making data generated by public bodies effectively sharable.
- TMX is dead (now definitely together with LISA).
- XLIFF natural successor (CNGL LRC Phoenix makes use of XLIFF as TM)

What Meatadata?

grammatical, syntactic, morphological, and lexical metadata that will facilitate Natural Language Processing (NLP), semantic, MT and other automated processing

- Content owners and transformers such as research institutes and universities (typical META-NET members) may have created advanced linguistic and/or semantic meta-data that might be of excellent use for MT technology and service providers.
- m4loc (Moses for Localization)
- CNGL | LRC LKR → Phoenix

What Meatadata?

process and quality (P&Q) metadata

- crucial for mutual automated communications between content publishers and localization service providers (LSPs)
- Raw MT output e.g. is not suitable for MT training
- P&Q metadata will allow for advanced conditional workflow automations
- In fact, large XLIFF implementers such as Oracle WPTG do use this faculty of XLIFF even now

What Meatadata?

tagging of culturally and/or legally targeted information

- The content authors and owners need to tell the localizers more than the ITS currently allows (just binary translate/do not translate, and there are at least three different possible XLIFF implementations)
- Legally targeted information needs other type of processing compared to culturally neutral description of a vacuum cleaner. Market specific safety regulations need different processing compared to culturally targeted marketing communication.
- This type of information will again allow for advanced conditional workflow automations.

Homes for LT standardization?

Leverage best practices of existing localization standards such as **OASIS XLIFF**, **LISA OSCAR** TBX, TMX, SRX and GMX

Leverage best practices of existing localization standards such as **OASIS XLIFF**, LISA OSCAR [**ISO TC37**] TBX, [**legacy**] TMX, [future **Unicode** successor standards of] SRX and GMX.

Furter develop **W3C ITS** and **RDF**. Create conscious standardized hooks for ITS and RDF in XLIFF.

Homes for LT standardization and their roles?

OASIS – home of the core standard XLIFF and the reference architecure OAXAL.. (UBL, ebXML, Translation Web Services)

W3C – home of ITS and RDF

Unicode – to form shortly an L10n TC. Initiative of Helena Shih Chapman from Wlatham, MA IBM office. Natural home for SRX and GMX successor standards

ISO TC37 – ISO not a good body for standards development, excellent for secondary publishing to secure governmental enforcement. After TBX and SRX, XLIFF goes this way..

The LT standards development within OASIS, W3C, and Unicode and secondary publishing in ISO TC37 must be coordinated and orchestrated.

Why XLIFF, and why 2.0?

- Uptake in industry adoption and community involvement last years
  - Roughly since SDL acquisition of Idiom (Feb 2008)
- XLIFF is the open standard bi-text format
  - Attractive for big publishers who want to go descriptive rather than prescriptive
- Extensibility – adoption driver and killer
  - **Very low common denominator**
  - Need for XLIFF 2.0 minimal and modular

What are the issues of 1.2?

- Reduced interoperability due to
  - Critical functionality in proprietary extensions
  - Semantic overload of key structural elements
  - Ecclectic approach to inline markup
  - Lack of conformance clause and processing expectations
- For all that, XLIFF 1.x is still a huge success!
  - Although the interoperabitilty is not plug&play it is still there..

XLIFF 1.2 GOs and NO GOs

http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html#AppTree

GOs

<file>, <skl><source><target>, <alt-trans>

No GOs

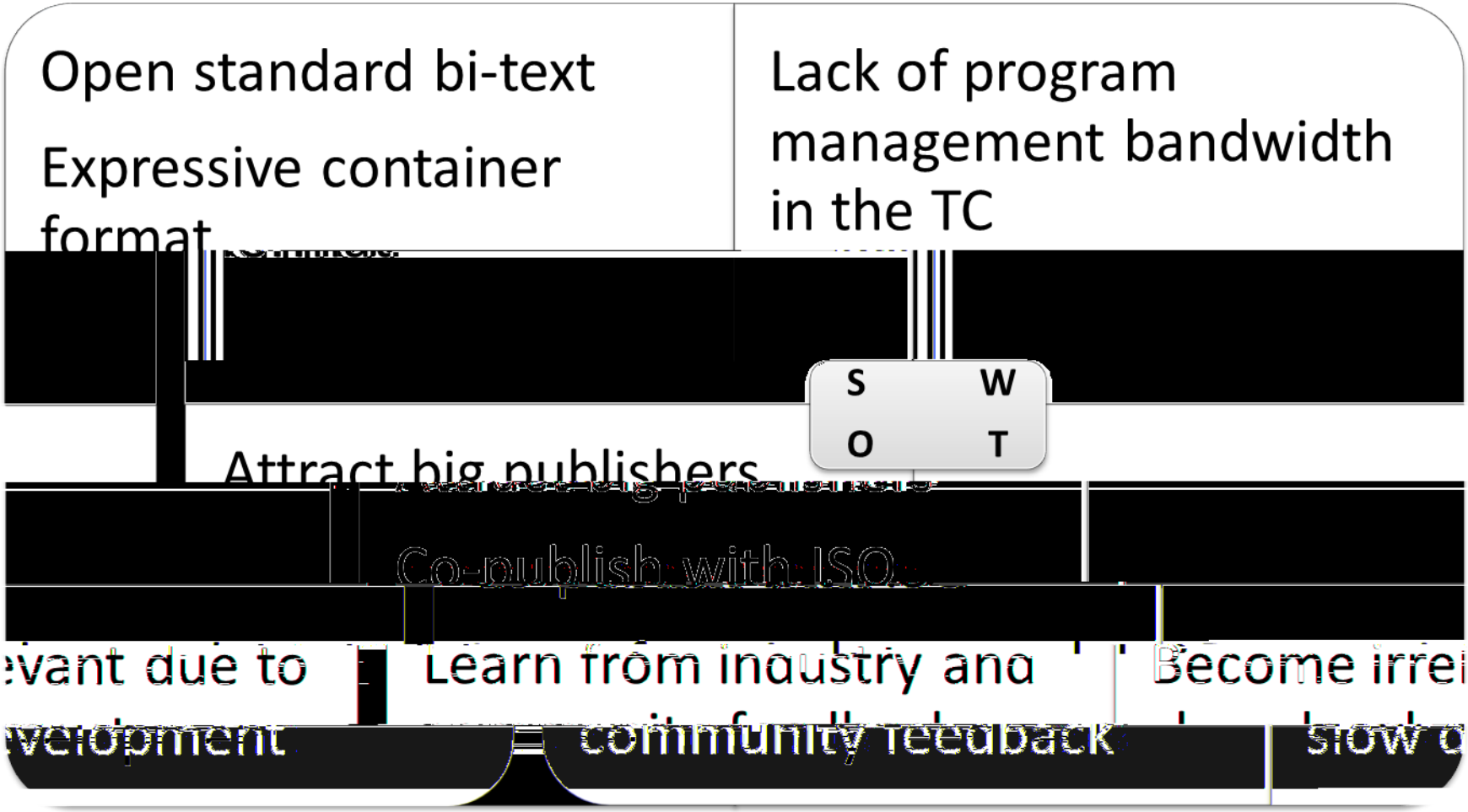
Generous extensibility, lack of conformance clause

Implementers ignoring 1.2 segmentation provision <seg-source><mrk mtype="seg" >

Mixed

<phase>, <group>, [inlines]

XLIFF 2.0 SWOT Analysis



Status of the SWOT Progress in 2011

- New manpower in the TC is likely to address the capacity issues
  - IBM rejoined the TC
  - Multicorpora and LIOX to send respresentatives
  - What about MS?
  - Inline Markup SC still needs more manpower and discussion with industry
  - DavidF from Moravia to LRC

Progress in 2011 continued

- Toolmakers willingly documenting their extensions and the semantics of their implementations
  - SDL, Kilgray, Multicorpora et.al.
  - TC prepares OASIS infra to display interoperability info on standing implementations
  - 2nd International XLIFF Symposium in Warsaw September 2011

XLIFF 2.0 SWOT Analysis

Persistent Strengths	Being well addressed by influx of new manpower. Toolmakers want to participate.
Good progress on collection of implementers' extension points, semantics etc.	In 2011 the TC should finish the initial requirements gathering and features definitions. Q12012 should see the new committee draft and Q2 the 2.0 standard

Challenges for 2.0

- Determine a powerful and compulsory core
  - Including processing requirements
  - Disambiguate core structural elements
- Sort out inline mark up salad
- Create meaningful extensions
- All that must happen in historically short and hence relevant time-frame
- Coordinate with W3C, Unicode and ISO TC37

Q&A at the end of the whole L10n session

Thanks for your attention!

david.filip@ul.ie