Make the Multilingual Web Work

TPAC 2012 Breakout Session

1. MULTILINGUALWEB-LT WORKING GROUP

Gap: Metadata in the “deep Web”

- Input from the data base – the “hidden web”:

  „Ob <term>Postbank direkt</term>,
  <term>Online-Banking</term>,
  <term>Online-Brokerage</term> …“

- Output on the Web:

  „Ob Postbank direkt,
  Online-Banking,
  Online-Brokerage …“

fixed terminology (= metadata) …

publication process

… is lost on the Web ☹

Filling the gaps: Internationalization Tag Set (ITS) 2.0

- Defining metadata (ITS 2.0 “data categories”) for language technology in the Web, e.g.
  - Machine translation
  - Localization workflows
  - Example: “translate” attribute in HTML5
- Where is the metadata needed:
  - In Web content, e.g. HTML5
  - In the “deep Web” (e.g. XML)
  - In RDF, see http://www.w3.org/TR/its20/#conversion-to-nif
  - In Localization related formats like XLIFF

Metadata example: “Translate” in HTML5 (=Web) and XLIFF (*one* deep Web format)

```
<xliff ...> ...
 <trans-unit id="1">
  <source xml:lang="en">The <mrk mtype="protected">World Wide Web Consortium</mrk> ...!</source>
  <target> ...
</xliff>
```

```
<!DOCTYPE html>
<html> ...
<p>The <span translate=no>World Wide Web Consortium</span> is making the World Web Web worldwide!</p>...</html>
```

Filling the gaps

- DFKI (coordinator)
- Trinity College Dublin
- Dublin City University
- Moravia
- Univ. of Econ. Prague
- Microsoft
- Enlaso
- Institut Jozef Stefan
- University of Limerick
- Cocomore
- Linguaserve
- VistaTEC
- Lucy Software
- Alchemy Software

Also: Adobe, Baidu, CNR, DERI, EMI, Inria, Opera, UPM, Vrije Universiteit

Eye catcher: list of ITS 2.0 data categories

- Translate, Localization Note, Terminology, Directionality, Ruby, Language Information, Elements Within Text, Domain, Disambiguation, Locale Filter, Translation Agent Provenance, Text Analysis Annotation, External Resource, Target Pointer, Id Value, Preserve Space, Localization Quality Issue, Localization Quality Précis, MT Confidence, Allowed Characters, Storage Size

Reference Implementations

- CMS – Localization chain (= XLIFF) integration
- Online MT systems
- Deep Web information and MT training

EXAMPLE USE CASES: SIMPLE MACHINE TRANSLATION

Simple Machine Translation

Description

- XML and HTML5 documents are translated using a machine translation system, such as Microsoft Translator.
- The documents are extracted based on their ITS properties and the extracted content is send to the translation server. The translated content is then merged back into its original XML or HTML5 format.

Data Categories

- Translate
- Locale Filter
- Element Within Text
- Preserve Space
- (Domain)

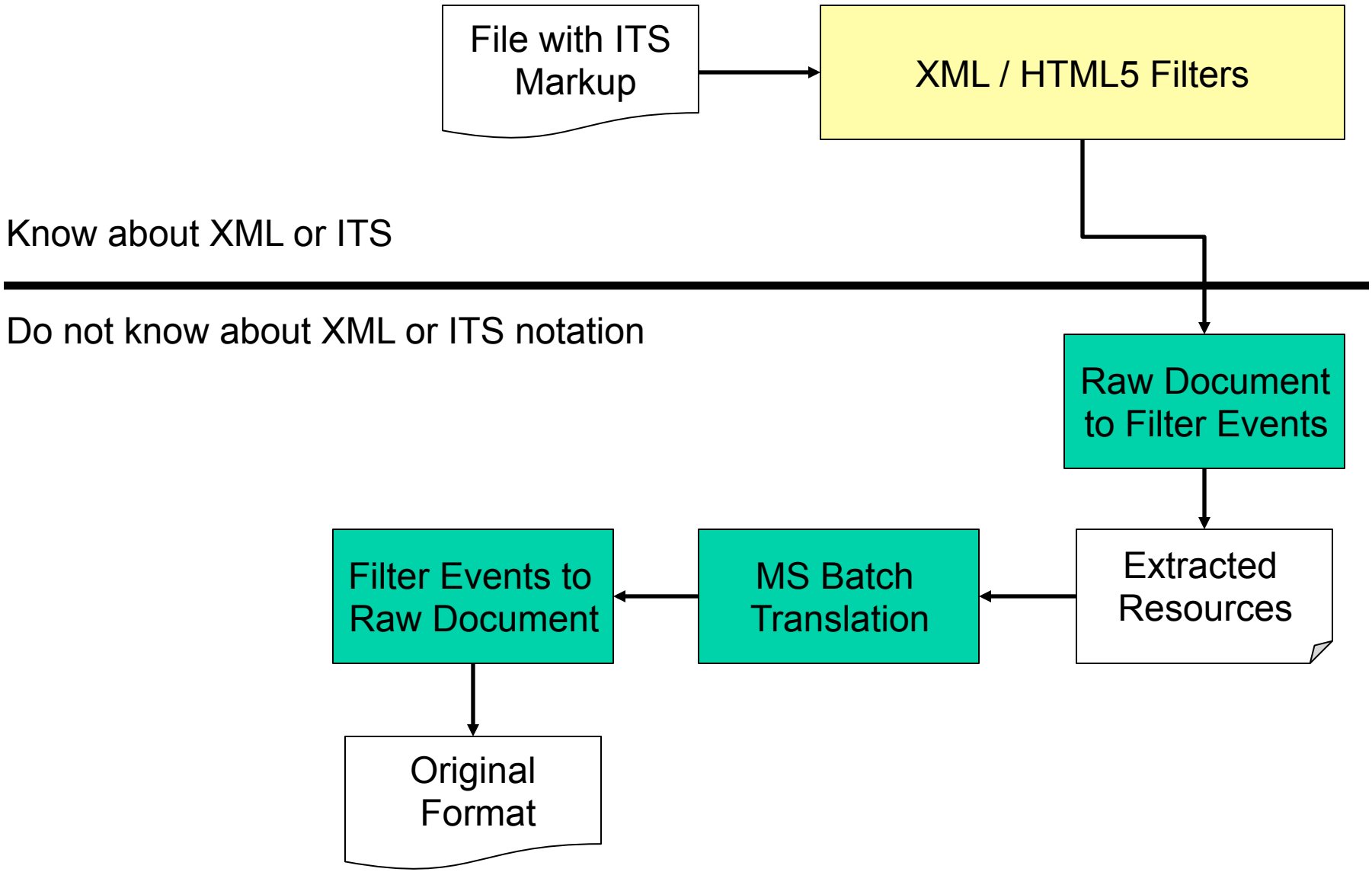
Benefits

- The ITS markup provides the key information that drives the extraction in both XML and HTML5.
- Information such as preserving white space can also be passed on to the extracted content and insure a better output.

Simple Machine Translation

- **Translate** - The non-translatable content is protected.
- **Locale Filter** - Only the parts in the scope of the locale filter are extracted, the others are treated as 'do not translate' content.
- **Element Within Text** - The information is used to decide what elements are extracted as in-line codes and sub-flows.
- **Preserve Space** - The information is passed on to the extracted text unit.
- **(Domain)** - The domain values are placed into a property that can be used to select an MT engine.

Simple Machine Translation



More Information about implementation

- Project wiki:
  http://www.opentag.com/okapi/wiki/
- Project source code:
  http://code.google.com/p/okapi/
- Continuous integration:
  https://okapi.ci.cloudbees.com/
- Maven repositories:
  http://repository-okapi.forge.cloudbees.com/release/
  http://repository-okapi.forge.cloudbees.com/snapshot/
- Developers mailing list:
  https://groups.google.com/group/okapi-devel/

Involving other communities

- ITS 2.0 implementers gathering & XML community reach out at XML Prague 2013
  - 8. February 2013, Prague
- MultilingualWeb workshop
  - 12-13 March 2013, Rome
  - Register at http://www.multilingualweb.eu/register
- More to come in 2013

Current state of MLW-LT WG

- ITS 2.0 moving to last call in November
- Feedback on http://www.w3.org/TR/its20/ needed now
  - Data categories
  - Usage in HTML5 or XML
  - Usage via conversion HTML5 > RFD, see http://www.w3.org/TR/its20/#conversion-to-nif
- When? For example, today ☺
- Or Thursday – Friday (better)
  - MLW-LT Working Group meeting

2. INVOLVING COMMUNITIES = “MULTILINGUALWEB” WORKSHOPS

MultilingualWeb
http://www.multilingualweb.eu/

- EC funded workshop series
- Broad topic “Multilingual Web”
  - Cross-community
  - Detecting gaps that hinder progress of multilingual Web
  - Bring stakeholders together that can close the gaps
- One outcome: forming of MLW-LT working group
  - Focusing on metadata gap
  - Creating reference implementations and doing standardization

Stakeholders

- Developers
  - E.g. browser implementers
- Creators
  - CMS central
- Localizers
  - Translation agencies / departments
- Machines
  - Machine translation, cross-lingual search, …
- Users
  - You☺
- Policy makers
  - E.g. governments

Outcome: huge community

- Information sharing, see http://www.multilingualweb.eu/en/documents
- Detecting issues and areas of interest – *small* subset, see also http://tinyurl.com/mlw-tcworld-2012
  - Support of user preferences
  - Harmonization of MultilingualWeb sites
  - Interop of implementations
  - Inter-language links
  - Too many standards in some areas, e.g. localization
  - Use of language technology

3. YOU

Questions

- What are your issues with the MultilingualWeb
  - In general
  - Specific to HTML5, e.g. internationalization issues related to bidirectional text, international layout
  - Specific to ITS 2.0
- What areas should the MultilingualWeb community work with more closely
  - E.g. Semantic Web?
- What synergies do you expect form EU or other funding?