Automating Multilingual Metadata Vocabularies

Alejandro Bia
y
Juan Malonda
Departamento de Estadística, Matemáticas e Informática
Universidad Miguel Hernández

Jaime Gómez
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Markup and meaning (2)

Sperberg-McQueen (2000):

- *"Markup has meaning."*
- *"…How does markup mean? Because means something, …we know certain things."*
- *"…the meaning of markup is the set of inferences it licences."*

C. M. Sperberg-McQueen, Claus Huitfeldt and Allen Renear, *Meaning and Interpretation of Markup not as simple as you think.*

Extreme Markup Languages, Montreal, 15 August 2000.

Markup and meaning (3)

- Understanding XML tags is key to correctly delimit text or metadata structures.
- The understanding is lost when tags names (elements, attributes and attribute values) are in a foreign language.

Example: `<название главный="тип">`

Previous work

WU, Pei-Chi (2000): "Translation of Multilingual Markup in XML", 2000 International Conference on the theories and practices of Electronic Commerce, Association of Taiwan Electronic Commerce, Taipei, Taiwan, October 2000

URL: http://www.atec.org.tw/ec2000/PDF/14.2.PDF

- In a couple of words: not much.
- However, we found an article from Pei-Chi Wu [6], that addresses the problem of translating a tagset from English to Chinese for easier understanding and more accurate markup.
- As this author states: "In Extensible Markup Language (XML), users can even define their own markups using local languages. These are widely accepted practices to make documents more easily grasped by local users".
- Wu's paper addresses the issue of multilingual markup, proposes a bilingual translation process, and discusses its potential applications to electronic commerce.

Our reasons for multilingual markup

- **MCDL** → multidisciplinary project
- proof-reading and markup team → largest group (40)
- This team's background → basically in Spanish Language and Literature.
- Highest time cost: markup learning and training
- Consequence → necessity of markup scheme in Spanish

Objective… (2)

OBJECTIVE:

Introduction of standard markup schemes (generally in English) in non-English-speaking communities.

(these are, in some cases, communities where XML for electronic publishing is still rare)

Automatic generation of markup translators: starting point

Defining the set of possible translations of:

- element names,
- attribute names,
- and attribute values

to the different target languages.

XML FILE

What is the best place to record multilingual names?

The requirements are:

- The structure must be easy to parse (XSLT friendly).
- It must be easy to associate an object name in one language with its translation to another language, for any pair of given languages.
- The notation used to specify the original name and its translations must be clear and easy to use for a human reader. Also to produce documentation.

Automatic generation of markup translators: DTD for XML translation mapping file

```
<!ELEMENT TAGMAP (ELEMENT)+ >
<!ELEMENT ELEMENT (ATTR)* >
   <!ATTLIST ELEMENT
         en CDATA #REQUIRED
         es CDATA #REQUIRED
         fr CDATA #REQUIRED>
<!ELEMENT ATTR (VALUE)* >
   <!ATTLIST ATTR
         en CDATA #REQUIRED
         es CDATA #REQUIRED
         fr CDATA #REQUIRED>
<!ELEMENT VALUE EMPTY >
   <!ATTLIST VALUE
         en CDATA #REQUIRED
         es CDATA #REQUIRED
         fr CDATA #REQUIRED >
```


Automatic generation of markup translators: XML translation mapping document

```
<TAGMAP>
 ...
 <ELEMENT en="body" es="cuerpo" fr="corps"> </ELEMENT>
 ...
 <ELEMENT en="div0" es="div0" fr="div0">
   <ATTR en="lang" es="lengua" fr="langue"> </ATTR>
   <ATTR en="type" es="tipo" fr="type">
    <VALUE en="news" es="noticias" fr="nouvelles"/>
    <VALUE en="suggestions" es="sugerencias" fr="suggestions"/>
    <VALUE en="biblnews" es="novedades" fr="publications"/>
   </ATTR>
 </ELEMENT>
 ...
</TAGMAP>
```

ELEMENT NAMES

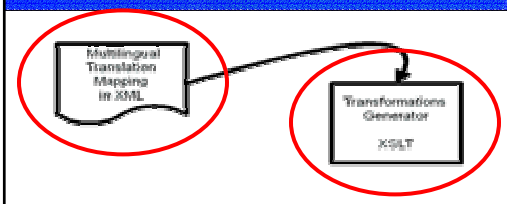
ATTRIBUTE NAMES

ATTRIBUTE VALUES

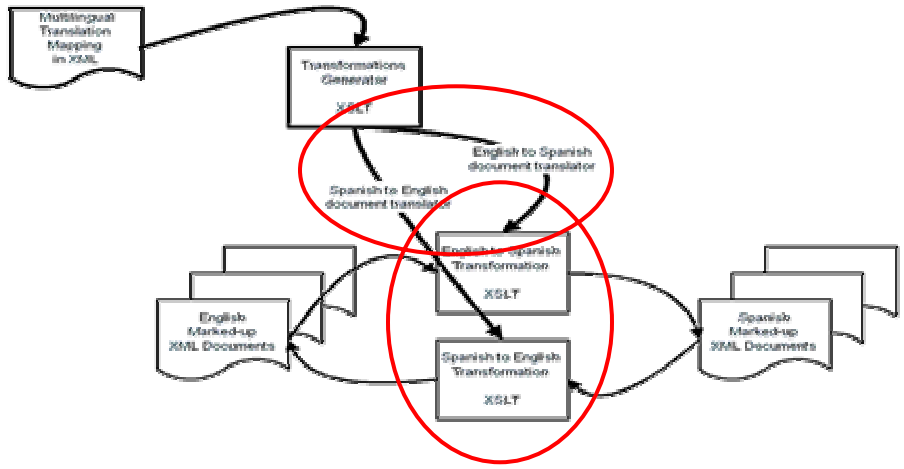
XML translation mapping for DC

```
<dcNames>
    <element ident = "title">
            <equiv lang = "es" value = "título"/>
            <equiv lang = "ca" value = "títol"/>
            <equiv lang = "fr" value = "titre"/>
            <equiv lang = "de" value = "Titel"/>
            <desc lang = "en">A name given to the resource.</desc>
            <desc lang = "es">Nombre dado al recurso.</desc>
    </element>

    <element ident = "creator">
            <equiv lang = "es" value = "creador"/>
            <equiv lang = "ca" value = "creador"/>
            <equiv lang = "fr" value = "créateur"/>
            <equiv lang = "de" value = "Ersteller"/>
            <desc lang = "en">An entity primarily responsible for making the content of
                              the resource.</desc>
            <desc lang = "es">Entidad principal responsable de hacer el contenido del
                              recurso.</desc>
    </element>

...
```


Translation of markup into Spanish

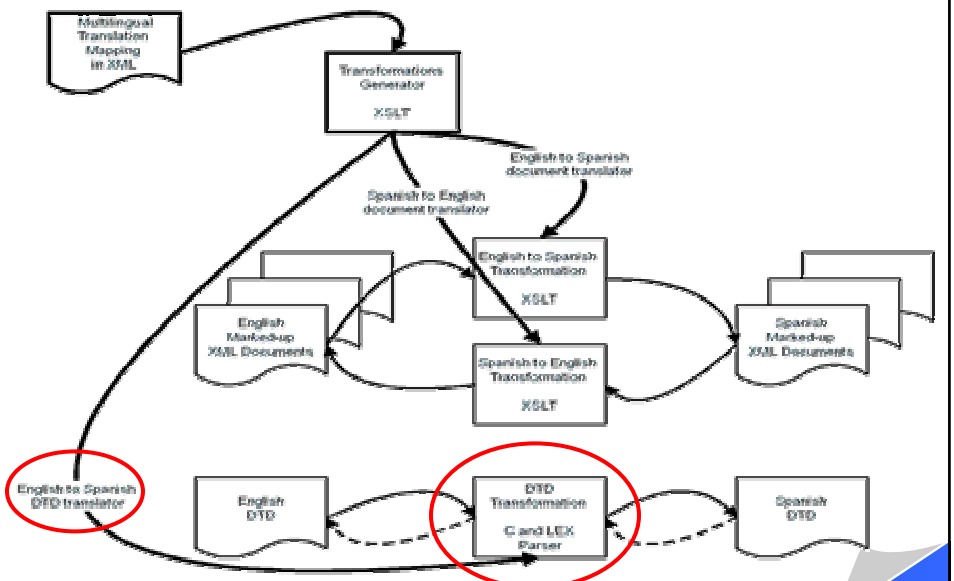


Translation of markup into Spanish



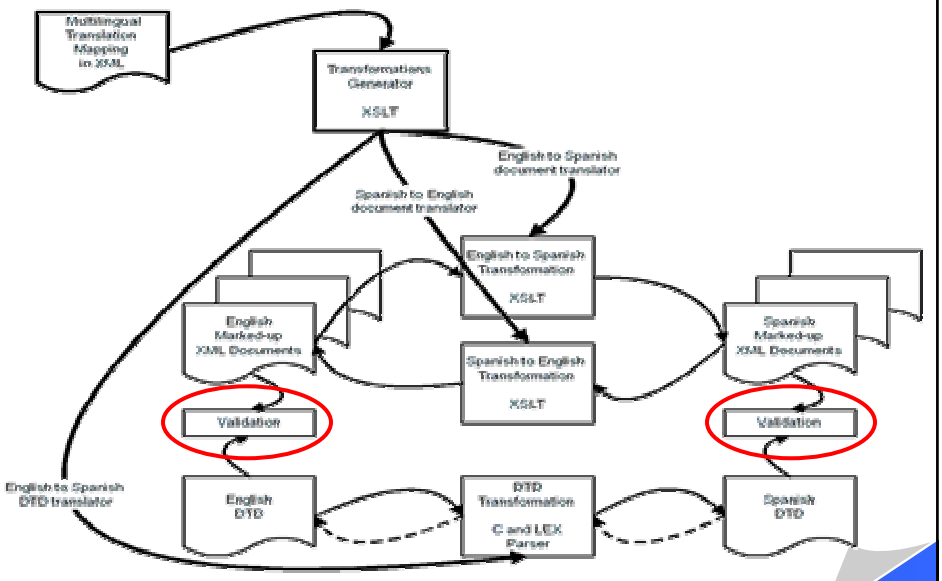
16

Translation of markup into Spanish



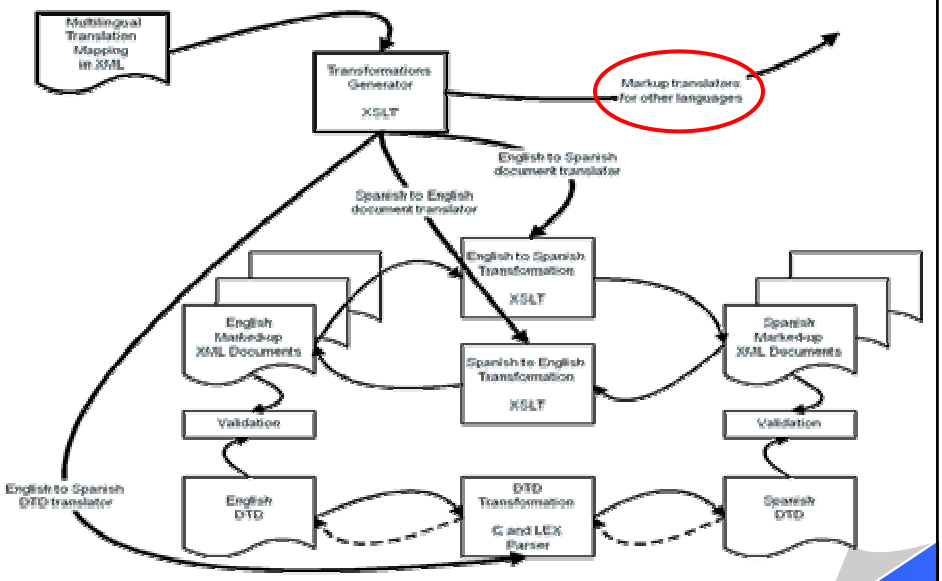
17

Translation of markup into Spanish



18

Translation of markup into other languages



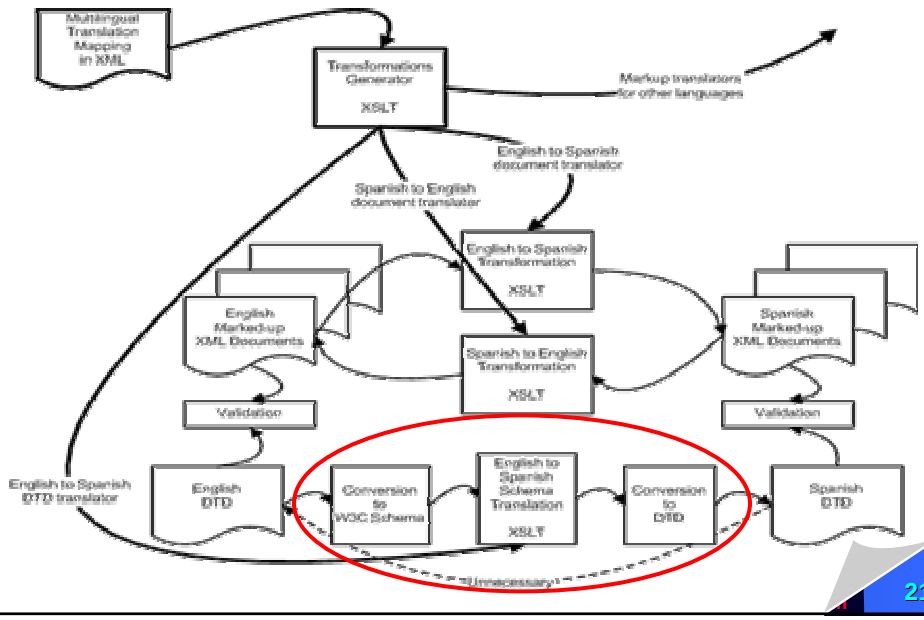
19

DTDs vs Schemas

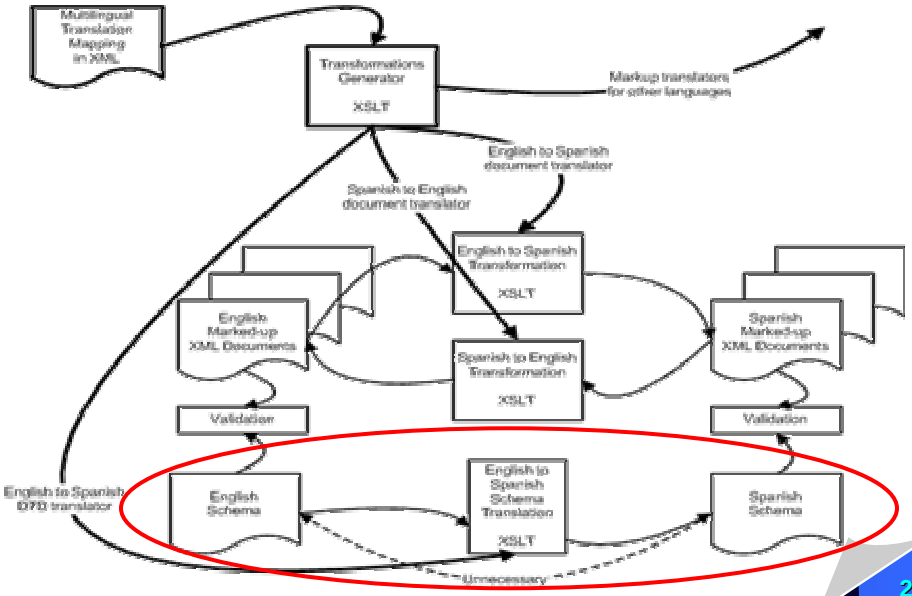
- DTDs are accepted by all XML/SGML editors
- DTDs are more compact

- Schemas are larger in size, but clearer to read.
- Schemas are easier to process and validate (they are XML)
- Schemas allow new features as data types

Translation of markup into Spanish



Translation of markup into Spanish



So what about stylesheets and other tools

Stylesheets and other tools need not be changed in any way.

The file can always be translated into English markup before processing.

This can be done in a fast and transparent way.

Now, let's see this work.

Advantages (1)

1. **Learning times** → noticeably reduced
2. **Production times** → also reduced
3. **Markup quality** → increased
4. **Using markup in one's own language**
   → meaning of markup is not lost

Advantages (2)

5. Having a standard vocabulary in one's own language may prevent developing custom vocabularies.
6. Spreading the use of standard markup vocabularies is good for document interchangeability.
7. Cooperative multilingual projects may benefit (interesting in the context of the EU)

Conclusions (1)

- We started by making a translation of TEI into Spanish.
- We ended up with a general set of tools to convert any markup vocabulary to many languages.

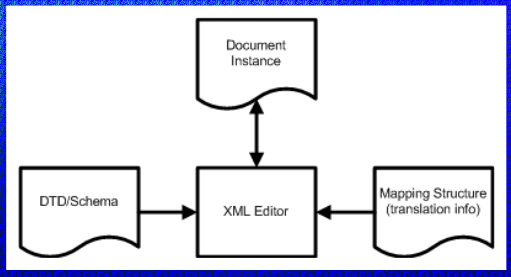
Conclusions (2)

- There may be better implementations to solve this problem.
- However, the problem is an interesting one and deserves to be solved.
- The development and use of multilingual tagsets should be spread to become common practice.
- Re-engineer tools in Java or C++, to provide faster performance and a nicer interface.

Future work

Embedding into editors

- Translation on-opening the document
- Changing views on the fly
- Apart from the document instance and the DTD/Schema for validation, a mapping structure with information for the translation is also required.



Future work

Within the I18N project of the TEI-C

- Add new languages.
- Implementation as a client-server Web service.
- Implement multilingual markup within the Roma DTD/Schema generator

Questions and comments...