On multilingual web sites

M.T. Carrasco Benitez
European Commission, Luxembourg, July 2012, version 1.0

1. Abstract

Multilingual Web Sites (MWS) refers to web sites that contain multilingual parallel texts; i.e., texts that are translations of each other. For example, most of the European Institutions sites are MWS, such as Europa [EU]. The main point of views are:

- **Users** should expect the same multilingual behaviour when using different browsers and/or visiting different web sites.

- **Webmasters** should be capable of creating quickly high quality, low cost MWS.

This is a position paper for the Dagstuhl Seminar on the Multilingual Semantic Web. Personal notes on this event are at:

http://dragoman.org/dagstuhl

2. Relevance

MWS are of great practical relevance as there are very important portals with many hits; also they are very complex and costly to create and maintain: Europa is in 23 languages and contains over 8 million pages. Facilitating and enjoying this common experience entails **standardisation**: current multilingual web sites are *applications* incompatible with each other. There is a Multilingual Web Sites Community Group at the W3C [CG].

3. Point of views

3.1. User

From a users point of view, the most common usage is **monolingual**, though a site might be multilingual; i.e., users are usually be interested in just one language of the several available at the server. The language selection is just a barrier to get the appropriate linguistic version. One has also to consider that some users might be really interested in several linguistic versions. It is vital to agree on common behaviours for users: browser-side (*language button*) and server-side (*language page*).

3.1. Webmaster

Webmaster refers to all the aspect of the construction of MWS: author, translator, etc. The objective is the creation of high quality low cost MWS. Many existing applications have some multilingual facilities and (stating the obvious) one should harvest the best techniques around.

Servers should expect the same application programming interface (API). The first API could be just a multilingual data structure. The absence of this data structure means that each application has to craft this facility; having the same data structure means that servers (or other programs) would know how to process this data structure directly. It is a case of production of multilingual parallel texts: the cycle *Authorship, Translation and Publication chain* (ATP-chain) [MPT].

4. Wider context

- **Language vs. non-language aspects**: differentiate between aspects that are language and non-language specific. For example, the API between CMS and web server is non-language specific and it should be addressed in a different forum.
- **Language as a dimension**: as in TCN, one should consider language a *dimension* and extend the concept to other areas such as linked data. Consider also *feature negotiations* as in TCN.
- **Linguistic versions**: the *speed* (available now or later) and translation *technique* (human or machine translation) should be considered in the same model.
- **Unification**: multilingual web is an exercise in unifying different traditions looking at the same object from different angles and different requirements. For example, the requirements for processing a few web pages are quite different from processing a multilingual corpus of several terabytes of data.

4. Multidiscipline map

- Web technology proper
    - Content management systems (CMS), related to authoring and publishing
    - Multilingual web site (MWS)
    - Linked data, a form of multilingual corpora and translation memories
    - Previous versions in time, a form of archiving [MEMENTO]
- Traditional natural language processing (NLP)
    - Multilingual corpora, a form of linked data [MUSET]
        - Source documents and tabular transformations, the same data in different presentations
    - Machine translation, for end users and prepossessing translators
- Translation
    - Computer-aided translation (CAT)
        - Preprocessing, from corpora, translation memories or machine translation
    - Computer-aided authoring, as a help to have better source text for translation
    - Localisation
    - Translation memories (TM) [TMX], related to corpora and linked data
- Industrial production of multilingual parallel publications
    - Integration of the Authorship, Translation and Publishing chain (ATP-chain)
    - Generation of multilingual publications
    - Official Journal of the European Union [OJ]

5. Disclaimer

This document represents only the views of the author and it does not necessarily represent the opinion of the European Commission.

6. References

[CG] Multilingual Web Sites Community Group; http://www.w3.org/community/mws
[EU] Europa; http://europa.eu
[MEMENTO] Memento - Adding Time to the Web; http://mementoweb.org
[MPT]Open architecture for multilingual parallel texts; http://arxiv.org/pdf/0808.3889
[MUSET] Multilingual Dataset Format; http://dragoman.org/muset
[PN] Personal notes for this event; http://dragoman.org/dagstuhl
[OJ] Official Journal of the European Union; http://publications.europa.eu/official/index_en.htm
[TCN] Transparent Content Negotiation in HTTP; http://tools.ietf.org/rfc/rfc2295.txt
[TMX] TMX 1.4b Specification; http://www.gala-global.org/oscarStandards/tmx/tmx14b.html