Metadata (and more ...) for the “ubiquitous multilingual web”

Felix Sasaki
Univ. of appl. Sciences Potsdam / W3C German-Austrian Office
What is the multilingual web?

• Requires proper Internationalization, like preparing your systems to deal with
  – Unicode
  – Bidirectional Text
  – Language specific layout
  – Etc.

• “Systems”: may be authoring system (CMS), web servers, (mobile) browsers etc.
How will Language Technology become part of the multilingual web?

• The same like “traditional” multilingual web: prepare authoring systems, servers, browsers etc. for LT!

• One step is “Metadata for the multilingual web”: ITS (Internationalization Tag Set) provides Metadata like
  – “Translate”: translatability of pieces of XML
  – “Terminology”: identification of terms
  – “Elements within Text”: Text flow properties
Applicability of ITS for LT – current state

• “Translate”: Tell MT-system what content (not) to process

• “Terminology”: easy connection to terminological data bases

• “Elements within Text”: configure processes like machine translation or spell checking (“This is a footnote inside a sentence, be careful to segment them properly!”)
My dream no. 1: ITS in more areas

• Currently focusing on XML
• “Data Categories” like “Translate” or “Terminology” have a wider scope
  – Different types of data: RDF / HTML / WORD / …
  – Different systems: Browser, authoring tools, CMS, …
My dream no. 2: (ITS as a) hook

• ... to bridge the gap between LT and the web infrastructure
• Have metadata about grammars, lexica, MT tools etc. in web standards like HTML
• Creating LT applications in CMS systems, on web servers, in browsers ... not via the web, but LT as part of the web!
The goal: reaching “the long tail” of a diverse multilingual society

- Earthquake scenario: A doctor who doesn’t understand a patient’s language can’t wait for a (human or sophisticated MT) translator

- Need to plug-in LT resources with his preferences (“doctor” in “country XYZ” with situation “patient care”) ASAP

- Availability of LT not on, but in the Web may help to build the applications from scratch as needed – “just in time Language Technology”
Who has to work together?

1. Language Resources developer / standards-org (ISO), e.g. profiles for formats like LMF on the Web
2. LT developers: implementing 1. and other metadata in their tools / making their tools part of HTML / DOM / CSS / HTTP / ... processing
3. Localization Industry: workflows to be applied in browsers / servers / etc. for “just in time” application of LT
4. Web Industry: implementing the hooks in HTML etc.
5. Societies (in Europe or e.g. India): providing requirements and feedback about “long tail” application scenarios
Where would this happen?

• Partially it is happening already
  – Within META-NET

• My dream no. 3: Both could build the path to a sustainable activity:

A “Language Technology” Working Group within W3C:
“LT not on, but as part of the Web”
Have we seen similar things?

• Yes no. 1: W3C Voice Browser Working Group
  – The place for speech technology on the Web
  – Research and appl. development meet

• Yes no.2: Semantic Web
  – Research about Semantics meet the Web

• Conclusion:
  It can work, we just have to try it!
Summary: “My dreams”

• Dream 1: Hooks for LT in the Web everywhere
• Dream 2: Hooks for all kinds of LT resources and applications / user needs
• Dream 3: An area where LT / Localization Industry / Web technology can meet – a W3C “Language Technology” Working Group
• A phrase for the goal:
  “ubiquitous multilingual web”