MLW-LT Call For Participation

David Filip
Dave Lewis
Felix Sasaki
Terminology

- CSA – Coordination and Support Action
- W3C – Worldwide Web Consortium
- WG – Working Group (in W3C)
- Deep Web, Surface Web
- LSP – Language Service Provider
- TM, MT, TMS
- CMS, CCMS
- OASIS DITA, XLIFF
LT-Web and MLW-LT

- LT-Web is an EC funded CSA
- LT-Web members will join W3C (or are members already) and will form the MLW-LT group
- All normative output of LT-Web will be generated solely through the MLW-LT WG
EC LT-Web and MLW-LT

LT-Web Consortium
General Assembly
One representative from every ben
Chaired by the CSAC

MLW-LT W3C WG
Governed by W3C
procedures and Charter

Project Coordination Team
(DFKI, UL, and TCD)

W3C WG Officers

Administrative and Financial Coordinator
(AFC)

CSA SA Coordinator (CSAC)

Workpackages 2 to 5
Who is in?

We want your logo here 😊
Standardization focus - Metadata

• Multilingual Web must be aware of linguistic and localisation processing
  – Process and Quality, Translatability, Legal, Terminology & Semantics..

• Three main in scope scenarios
  – Deep Web <-> LSP
  – Surface Web <-> Real Time MT
  – Deep Web <-> MT Training
  – All other scenarios are out of scope

• Reference implementations, XLIFF roundtrip prototypes, and test suits for all three
Deep Web <-> LSP

- Deep Web is mostly XML and is being managed by CMS, ideally CCMS.
- Cocomore is involved in Drupal and Sharepoint based CMS and CCMS solutions.
- Passing process, terminology, and translatability metadata from CCMS onto down stream localisation chain actors.
Surface Web <-> Real Time MT

• Ensure that relevant Deep Web metadata will resurface in the rendered HTML, so that real time MT services can make use of them to improve their output

• Again, translatability or terminology metadata will be passed onto MT to improve results
Deep Web <-> MT Training

• Improve MT training through passing domain and processing related metadata
• This will allow for rapid creation of relevant training corpora, excluding upfront out-of-domain content, raw MT output etc.
Metadata

• "data categories" based on "W3C Internationalization Tag Set 1.0" relevant for the three scenarios:
  – Translate, Localization Note, Terminology, Language Information

• Further data categories:
  – Translation provenance, human post-editing, QA provenance, legal metadata, topic / domain information

• Everything is currently under consideration – your input counts!
Approach and Methodology

• Open Standard within W3C Internationalization Activity:
  – Transparent & Royalty Free

• Normative Processing Requirements
  – Based on in scope process models
  – Methodology how to expand to
    • Create conformant extensions
    • Enable future development

• Robust roundtrip implementations and test suits
  – bias for open source

• Close collaboration with OASIS XLIFF TC
Open Question(s)

• Breadth or Depth?
  – Scope? Too broad? Too Narrow? Additions?
  – Generalized Process Models as base for Normative Processing Requirements?
    Vs.
  – Define only data categories and give non-normative advice on processing?
  – More user scenarios?
  – Missed a critical category?