Linked Data Reuse in the Language Services Industry

David Lewis, Leroy Finn, Dominic Jones
Language & Localisation: Meta-Data and Interoperability

Localisation Industry

Translations + Term

Language Resource Community

Language Technology Community

Multilingual Web

OASIS

ISO

META-SHARE

Multilingual Web-LT

W3C
Language & Localisation: Meta-Data and Interoperability
Interoperability roadblocks in high variety:
- Content Management Systems
- Content formats

Increasingly dynamic

Add language resource curation
- Data driven MT & text analytics
Internationalisation Tag Set (ITS)

- Allows Localisation tools to be instructed to treat specific text in specific ways
- Annotates text in XML and HTML5 documents
- RDF Mapping – NLP Interchange Format
- Principles:
  - Minimise disturbance of original content
  - Don’t reinvent wheel
  - Link to existing meta-data before adding new
- Defined distinct, independent Data Categories
- ITS2.0 is “Dublin Core” of the Multilingual Web
<para>
Press the<br/>START button to sound the horn. The<br/>MAKE-READY/ RUN indicator flashes.
</para>

<para>
Press the<br/>START button to sound the horn. The<br/>MAKE-READY/ RUN indicator flashes.
</para>

<its:rules ... its:version="1.0">
   <its:translateRule selector="/uitext" translate="no"/>
</its:rules>
Here Content is the Data
So we must annotate textual content
- RDFa and microdata better for annotating elements

C. Lieske, F. Sasaki 2010
Process Monitoring in RDF

- Better up-front and real-time business decision support:
  - Use historic and aggregated metadata for inference
  - Content carries its own “audit trail”
- Metadata can be transformed to linked semantic data allowing:
  - Inferential queries across data siloes
  - New forms of data visualization
  - Extensible data relationships
Provenance Data: RDF-based logging

- W3C Provenance WG
  - [http://www.w3.org/2011/prov/](http://www.w3.org/2011/prov/)

ITS related entity subclass:
document, segment, analysed-text, term, translation, translation-revision

From: [http://www.w3.org/TR/prov-primer/](http://www.w3.org/TR/prov-primer/)
CMS to Localisation Workflow Roundtrip

Source CMS

Target CMS

QA viewer

CMSLION
Parse, filter, segment

XLIFF/PROV-O

RDF provenance store

Web-based Posteditor

Named Entity Recogniser – Enrycher

Workflow Management

XLIFF store

MT - Matrex

MT - Bing

MT – M4LOC

LKR

CAT

ITS +HTML5 +CMIS

ITS +SPARQL

ITS +XLIFF

ITS +XLIFF
ITS

Extract

Domain

Translate

Terminology

Post Edit

Quality Check

HTML5 CMIS

Content I18N

Curate Corpora

Workflow Analysis

Low Quality Issues

RDF: PROV-O + NIF

Text Analysis

HTML5

Machine Translate

OASIS

XLIFF

Segment

Publish Content

www.cngl.ie

Trinity College Dublin

DCU

Dublin City University

University College Dublin

University of Limerick

NDP

SFI
Active Curation

- Localisation Industry:
  - €33bn global industry 12% growth
  - Established commercial models for translation and term reuse
  - High interoperability costs
  - Long tail industry – 99% SMEs
  - SMEs struggle to benefit from leveraging reuse
  - Challenged to adapt machine translation and text analytics technology

- **Active Curation**: *On-demand, quality-driven* assembly of training corpora
  - Collect human linguistic judgements from across the workflow
  - Professional translator/reviewer, crowds, end user
  - Already shown 25% MT improvement via in-job retraining

- **Linked Language and Localisation Data**
  - Extending existing tools to generate process provenance data
  - Pooling/Bokering of data across value chains
Where Next?

• Help EC and other public bodies to bootstrap data
  – per year EC produces 11 million pages in 22 language => 14 billion triples
• Attribution/license annotation and access control – Controlled Linked Data
• Increase industrial engagement
  – Media localisation – audio, video, images
  – Non-localisation MT and TA applications
  – Multilingual content personalisation
  – YOUR USE CASE HERE
• Join new W3C Community Group:
  **Multilingual Linked Open Data**
  [http://www.w3.org/community/bpmlod/](http://www.w3.org/community/bpmlod/)
Thank You