The Localization Web: From L10n workflows to Linked Data

Dave Lewis
CNGL at Trinity College Dublin
Problem

• Multilingual web pages could offer an important language resource,
  – e.g. as parallel text for machine translation engine or multilingual term extraction

• Difficult to leverage, HTML is a publication format, it hides valuable translation info:
  – Translated sentence alignment
  – Term meta-data
  – Translation provenance: was it machine translated, transcreated, quality checked?

• Barrier to leverage by industry’s long tail of SME LSPs and clients
The Localization Web

• W3C Semantic Web standards allow data to be published on Web
  – Fine-grained URI-based inter-linking
  – Extensible meta-data
  – Standard Query APIs

• Enables a Localization Web
  – Terms and translations become linkable resources
  – Meta-data from L10n workflows adds value
  – Leverage in training Machine Translation and Text Analytics

The Localization Web = Decentralised Annotated Global Translation Memory and Term Base
LD allows linguistic data integration

- **singular**
- **plural**
- **[REDES]**

- **Form**
- **Red**
  - **femenine**
  - **“red”**

- **Wikilengua del español**

- **Sense**
  - **written form**
  - **translation es - en**
  - **“red”**
  - **“network”**
  - **written form**

- **Image**
- **Wikimedia**
- **Wikipedia**
- **“malla”**
- **“red”**
BabelNet

Traditional lexical resources

- fully-structured
- manually curated by experts
- available for a few languages
- difficult to maintain and update

Collaborative lexical resources

- semi-structured
- collaboratively built by the crowd
- highly multilingual
- up-to-date
Barak Obama if the 44th president of the United State of America. He was first elected in 2009.

Barak Obama si el 44 º presidente de los Estados Unidos de América. Ha fue electo primera vez en 2009.

http://www.ex.org/obama_en.html
http://www.ex.org/obama_es.html

Text: “Barak Obama if the 44th president of the United State of America.”
Lang:en

Derived From

Translated From

http://data.ex.org/String_0001

Term: “United State of America.”
Lang:en

Derived From

Translated By: Google Translate

http://data.ex.org/String_0002

Term: “Estados Unidos de América.”
Lang:es

http://babelnet.org/345621

Topic: Barak Obama
Lang: en
BirthDate: 1961-08-04
Spouse: Michelle Obama
Residence: White House

http://Dbpedia.org/Page/Barak_Obama

http://babelnet.org/57835

The Web of Content

The Localization Web

Translation Data

Terminology Data

Encyclopaedic Data
Use Cases

• Source Internationalisation
  – Term disambiguation
  – Term extraction with translation discovery
  – Auto-tag named entities with encyclopaedic reference for authors and translators

• Machine Translation
  – Consistent machine translation of terms
  – Pooling and discovery of parallel text for training

• Translation and Post-editing
  – Term definitions from open encyclopaedic data
  – Concordancing over a global TM
Data Management Lifecycles

Lex-concept lifecycle

Content lifecycle

Bitext lifecycle

Discover data

Revise and annotate

Discover & use

Create

I18n & source QA

Correct & refine

Trans QA

Post-edit

Automated translation

Discover & use

(Re)train -MT

Correct & refine

Publish

Publish

Publish

Publish

Correct & refine

Correct & refine

Consume
Realisable Benefits?

- How can Language Resource Publishers audit links to and use of resources & track ROI?
- Can Tool Vendors and Integrators expand markets with more open asset management offerings?
- How can SME LSPs gain resource sharing and pooling opportunities and avoid lock-in?
- Can LSPs and clients use Active Curation to quickly train domain specific SMT and text analytics components?
Data Management Needs?

• Assert ownership and attribution, licensing, access control?
• Persistent URLs?
• Open royalty free standards?
• Indexing is key, federated vs aggregated data?
• Third party submission of errors, QA, corrections? Publish action on submissions?
Data Access Needs?

• Query bitext on: languages, terms, MT engine, MT confidence, QA, translator qualifications, postedit characteristics?

• Query term base on: language, domains, context, semantic relations, provenance of lexical/conceptual data?

• Web API:
  – HTTP content negotiation (unicode extensions for translation?),
  – Format: RDF, TMX, TBX, RDF, JSON-LD,
  – SPARQL?
More Information

• Contact: dave.lewis@cs.tcd.ie
• http://www.falcon-project.eu

• See also:
  – Linked Data for Language Technology (LD4LT) W3C Community Group
  – http://www.w3.org/community/ld4lt/