Building the Localization Web

Dave Lewis
CNGL at Trinity College Dublin
• Localization Industry
• Data = Words (translations and terms)
• Exchanged in siloed value chains
• Statistical Language Technology improves cross-silo leverage
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
Use Cases

• Source Internationalisation
  – 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
Benefits

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

- Provide an **Open Schema** and **Integrated SaaS platform** for pooling and leveraging language resources and meta-data as linked data
- Enable **controlled, decentralised sharing** of resources and stand-off value-add annotation
  - Term or named entity annotation
  - Translation process provenance and QA
- **Active Curation** of resources and value add meta-data
- Monitor L10n workflows **end-to-end**
- Assemble corpora for domain-specific LT training **on demand**
Barak Obama if the 44th president of the United State of America. He was first elected in 2009.

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

Translation Data
- Text: “Barak Obama if the 44th president of the United State of America.”
  Lang: en
  Derived From: http://www.ex.org/obama_en.html
  Translated From: http://data.ex.org/String_0001

Terminology Data
- Term: “United State of America.”
  Lang: en
  Translation Of: http://babelnet.org/345621

Encyclopaedic Data
- Topic: Barak Obama
  Lang: en
  BirthDate: 1961-08-04
  Spouse: Michelle Obama
  Residence: White House
  Derived From: http://dbpedia.org/Page/Barak_Obama
  Translated By: Google Translate
  Translated From: http://data.ex.org/String_0002
  Derived From: http://data.ex.org/String_0002
  Terminology Data: http://babelnet.org/345621
  Terminology Data: http://babelnet.org/57835
Consortium

- Trinity College Dublin (IE)
  - L10n Interoperability (ITS2.0)
  - Linked Data Mapping and Link Quality
  - Federated Access Control
- XTM International (UK)
  - CAT/L10n management vendor and interoperability
- Interverbum Technology (SE)
  - Terminology Management
- Dublin City University (IE)
  - SMT and text analytics
- SKAWA Innovation (HU)
  - Web site translation (EasyLing), crowdsourcing
Seeking Collaborators

- Localisation Clients
- Language Service Providers
- Translators
- Language Resource Curators
- LR, LT and LD Researchers
- Standards Bodies
  - W3C (ITS), OASIS (XLIFF), ETSI, ULI:
  - FEIGILTTT workshop at LocWorld
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/