FALCON

# Building the Localization Web

# Localization, Data and the Web

- Disruptive Power of the Web:
  - Decentralised publishing
  - Hyperlinks to recommend and attribute resources enables global search
  - Now works with data as well as content
- Localization Industry:
  - Data = Words (translations and terms)
  - Exchanged in siloed value chains
  - Statistical Language Technology improves cross-silo leverage

# Problem

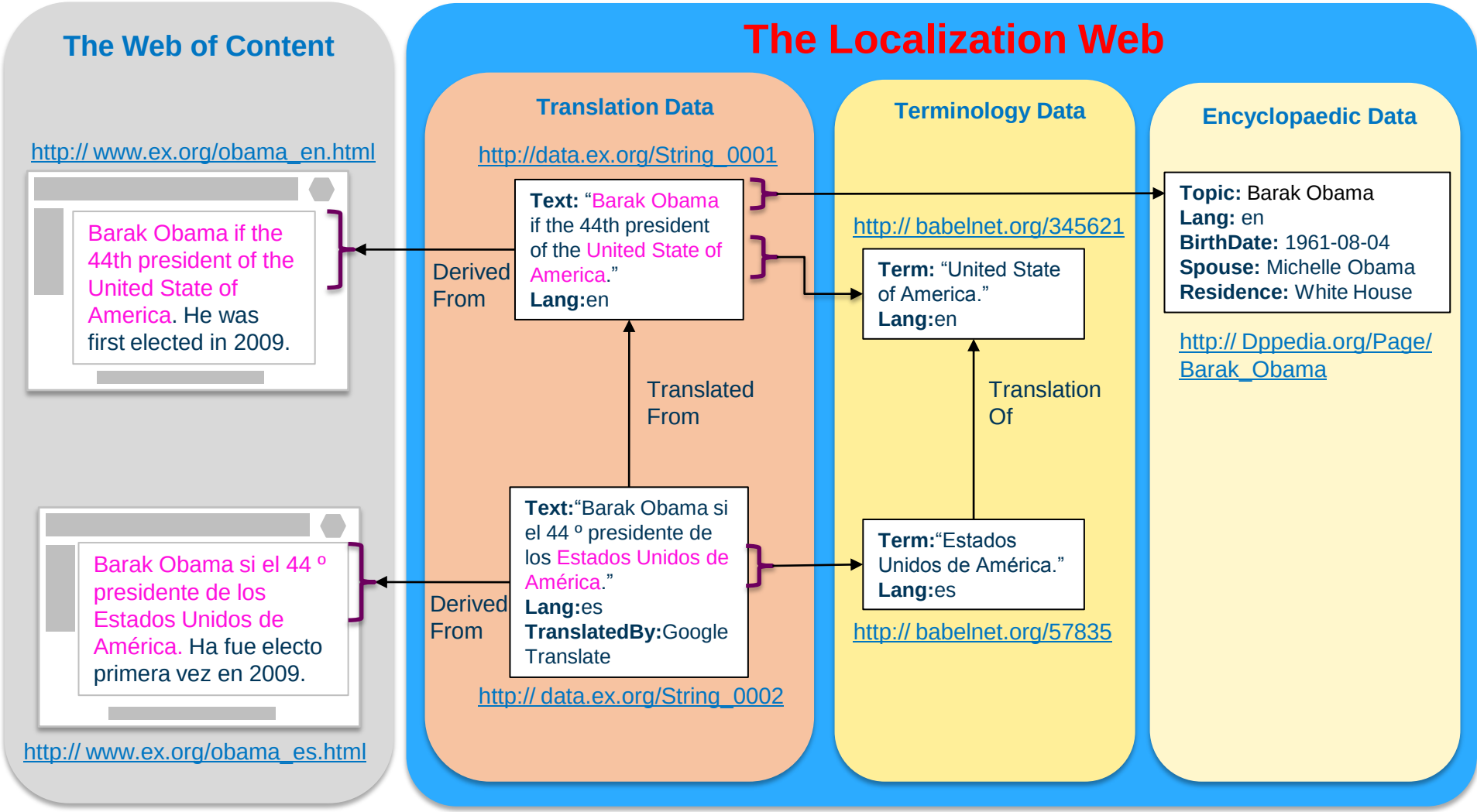
- Multilingual web pages could offers an important language resource,
  - e.g. as parallel text for machine translation engine or multilingual term extraction
- Difficult to leverage, HTML is a publication format, it hides valuable translation info:
  - Translated sentence alignment
  - Term meta-data
  - 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**

# Words as Resources on the Web



# 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

# 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

# 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 that avoid lock-in
- LSPs and clients can use Active Curation to quickly train domain specific SMT and text analytics components

# 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
  - Pool TM and Termbases with content consumers and content partners
  - Improve data management of TM/TBs: cleaning, updating, retrieval
  - Expose content annotation to improve multilingual indexing, SEO
  - Especially for government bodies already operating under open data policies

# Seeking Collaborators

- Language Service Providers
  - Sharing/exchange model of TMs/TBs for small, resource-poor providers
  - Improved language resource management (cleaning, filtering, selection) yields better targeted MT and term extraction engines
  - Improved monitoring of post-editing and review productivity

# Seeking Collaborators

- Translators
  - Guidelines on sharing their work as linked open data
  - Contractual and copyright restrictions
  - Pool resources with peers
  - Grow your own MT

# Seeking Collaborators

- Language Resource Curators
  - Publishing and interlinking resources as linked data, e.g. META-SHARE linked data schema
  - Rich meta-data for translation memories and term bases
  - Publishing and maintaining resources directly from industrial localisation workflows

# Seeking Collaborators

- Language Technology Researchers
  - Ongoing access to translation and term annotation resources
  - Provenance meta-data offers better control in selecting training data
- Linked Data Researchers
  - Address data confidentiality and access control issues from L10n value chains
  - Text specific content annotation

# Seeking Collaborations

- International Standards
  - W3C Internationalisation and Data Activities
  - W3C Community Groups
    - Best Practice in Multilingual Linked Open Data
    - Linked Data for Language Technology
    - Onto-Lex
    - Open Annotation
  - OASIS:
    - XLIFF and OAXAL
  - GALA CRISP

# More Information

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