FALCON

Building the Localization Web

Dave Lewis
CNGL at Trinity College Dublin

CENTRE FOR GLOBAL INTELLIGENT CONTENT

XTM International

interverbum Tech

easyling
1-click website translation

Localization, Data and the Web

- 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

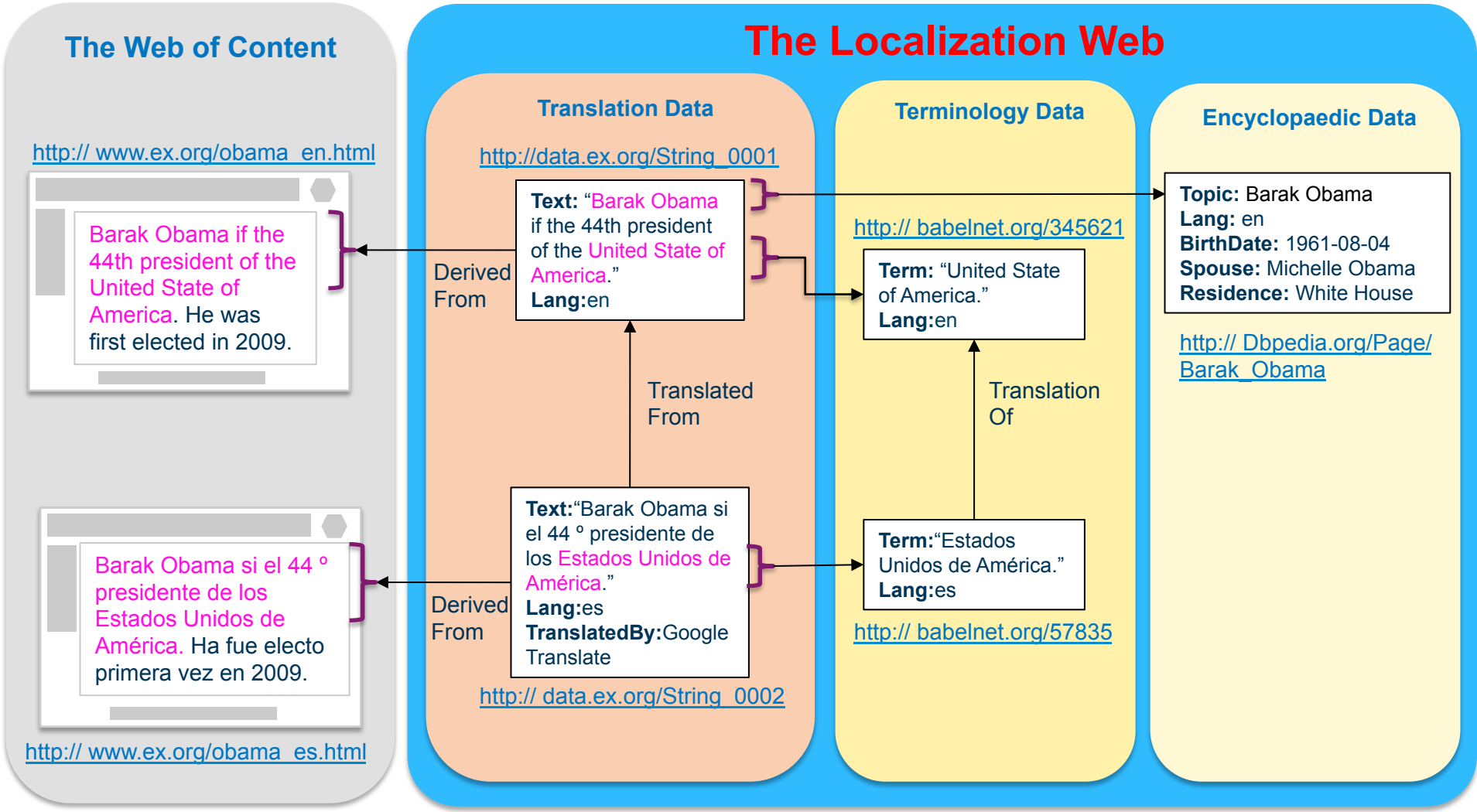
Benefits

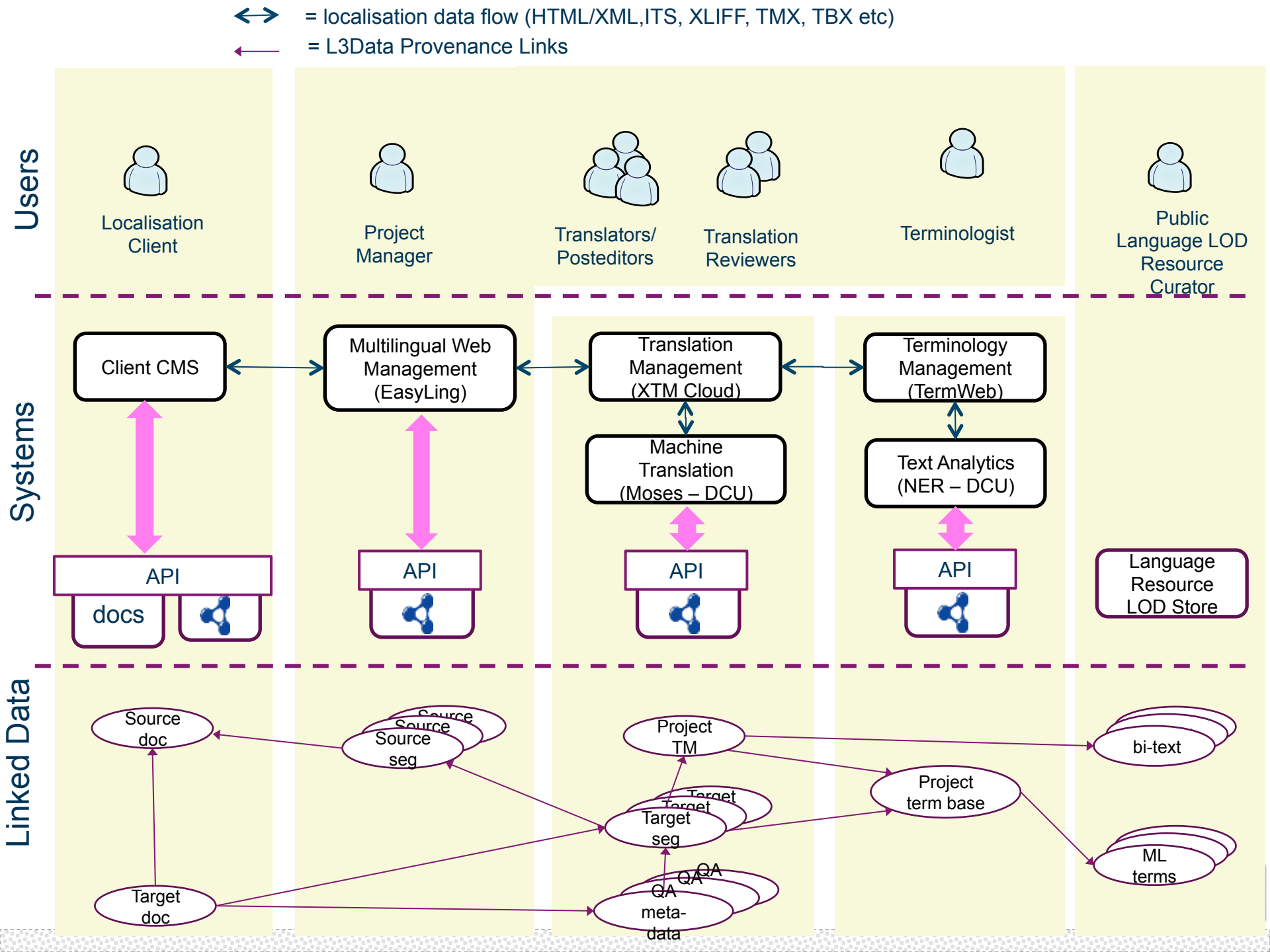
- Language Resource Publishers can audit links to and use of resources & track ROI
- Tool Vendors and Integrators expand markets with more open asset management offerings
- SME LSPs gain resource sharing and pooling opportunities and avoid lock-in
- LSPs and clients can use Active Curation to quickly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

Words as Resources on the Web




= localisation data flow (HTML/XML,ITS, XLIFF, TMX, TBX etc)
= L3Data Provenance Links
Users
Localisation Client
Project Manager
Translators/ Posteditors
Translation Reviewers
Terminologist
Public Language LOD Resource Curator
Systems
Client CMS
Multilingual Web Management (EasyLing)
Translation Management (XTM Cloud)
Terminology Management (TermWeb)
Machine Translation (Moses – DCU)
Text Analytics (NER – DCU)
API
docs
API
API
API
Language Resource LOD Store
Linked Data
Source doc
Source seg
Project TM
bi-text
Project term base
Target seg
Target doc
QA meta-data
ML terms

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:
  - FEIGILTT 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/