



Federated Active Linguistic data CuratiON FALCON

LT-Innovate Summit, 27th June 2013

Dave Lewis
CNGL - TCD

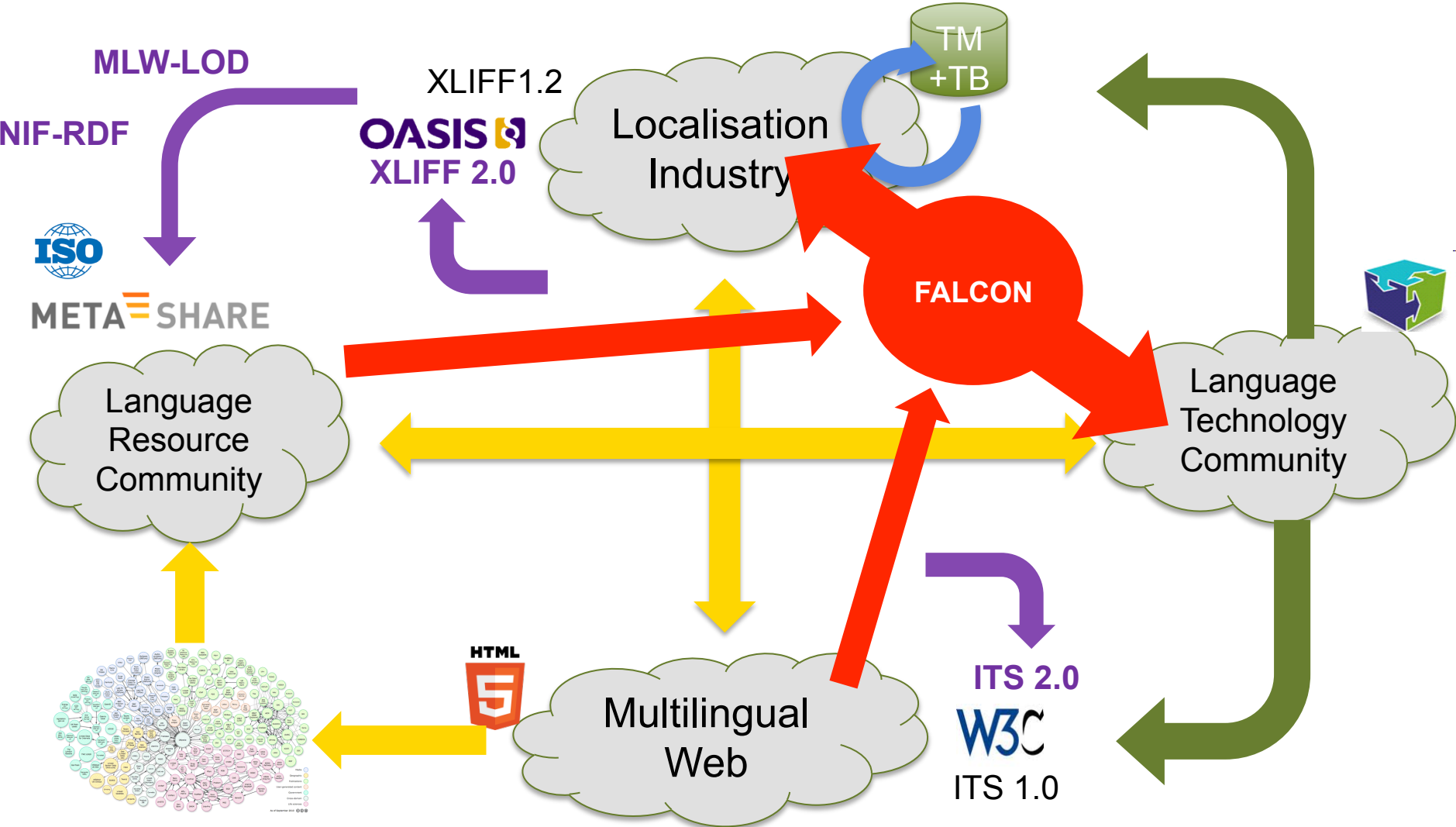


Aims of FALCON

Leverage the power of Linked Data for the Long Tail of the Localisation Industry

- L10n is a Big Data Industry
- Large-scale, Monetised Reuse of Translation Memories and Term Bases
- SMT and Text Analytics now also leverage these resources
- Large clients and LSPs curate and add-value to such resources as assets
- We aim to extend these benefits to SMEs

Where does FALCON sit?



FALCON Consortium



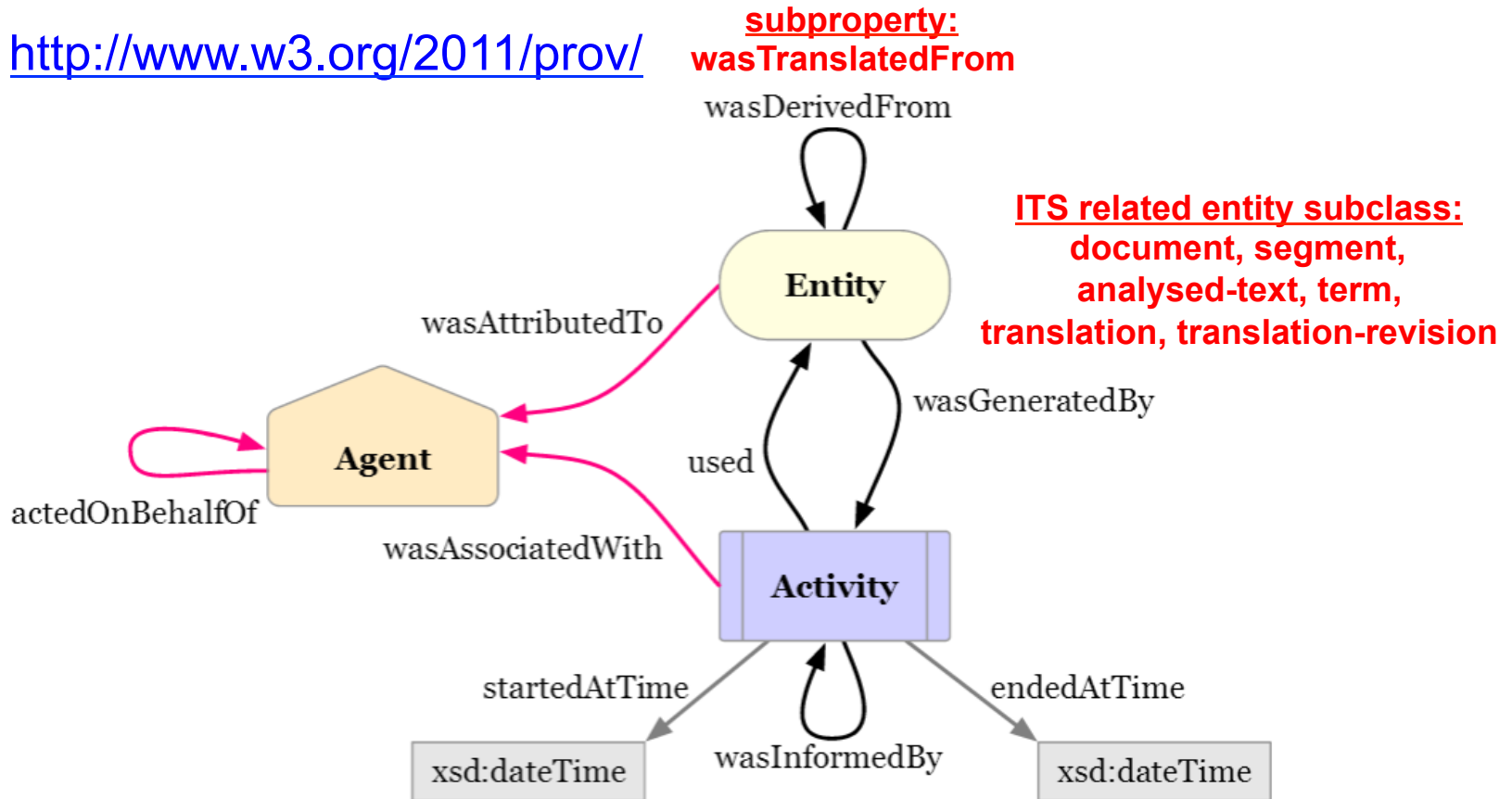
- Trinity College Dublin (IE)
 - LOD Mapping and Link Quality
 - Federated Access Control
 - L10n Interoperability
- XTM International (UK)
 - CAT/L10n management vendor and interoperability
- Interverbum Technology (SE)
 - Terminology Management
- Dublin City University (IE)
 - SMT and text analytics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

FALCON Approach

- Provide an Open Schema and SaaS platform for exposing language resources 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
- End-to-end process management
- On-demand assembly of domain specific LT training corpora

Provenance-Oriented Approach

- W3C Provenance WG
 - <http://www.w3.org/2011/prov/>



From: <http://www.w3.org/TR/prov-primer/>

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)


L3Data API




L3Data API



L3Data API



L3Data API



Language Resource LOD Store

Linked Data

Source doc

Source Source seg

Project TM

bi-text

Target doc

Target Target seg

Project term base

ML terms

QA meta-data

Value Network of Benefits

- Language Resource Publishers can audit links used in building other 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

Collaborate

- Seeking further collaborators:
 - Public bodies looking for more value-add from publishing language resource
 - Integrating with open source SMT and TA platforms
 - Standards and best practice in publishing language resources as linked data
 - Localisation clients or crowdsource communities interested in acting as trial users
- Projected start in Oct
- Contact: dave.lewis@cs.tcd.ie