



## FALCON: Federated Active Linguistic data CuratiON

### Building the Localization Web

**EU FP7  
SME STREP  
ICT SME-DCA Call 2013, FP7-ICT-2013-SME-DCA  
No. 610875  
<http://www.falcon-project.eu>**

<b>List of partners</b>
Trinity College Dublin, Ireland
XTM International, UK
Interverbum Technology, Sweden
Dublin City University, Ireland
SKAWA Innovation, Hungary

**Project duration: Oct 2013 — Sept 2015**

#### **Summary**

FALCON assembles a state of the art online translation tool chain that combines web site translation, translation management, computer aided translation and terminology management products. This tool chain tool chain has been enhanced with open-source automatic term extraction and machine translation technology. Iterative quality improvement in this language technology is delivered by using linked data to actively manage the curation and reuse of language resources within customer projects. The work demonstrates the integration the management of iterative SMT training with active curation of MT corrections and target term capture by post-editors in a live localisation workflow using this commercial tool chain. Key capabilities offered are:

- Targeting of translator effort to rapidly bootstrap automation quality; Translator judgements are the source of all resources used in machine translation and multilingual text analytics. Active curation targets available translator skills to optimise the improvement of language technology quality through incremental training. Specifically, the order of segment post-editing is optimised to harvest and integrate corrections to poor MT and translation of important terms as early as possible in a project.
- Building MT-ready terminology: Morphologically-rich multilingual terminology resources can dramatically improve MT quality through forced decoding of terms in source segments to their known translations
- Open Data for the analysis and reuse of translations and terms; Analysis of how available language resources can be applied to streams of new content will benefit from open data, provenance and licensing information formats to facilitate the discovery, selection, negotiation and reuse of resources, including those continuously generated by new translation projects and released through the public sector.