



Building the Localization Web

The Challenge: The disruptive power of the World Wide Web is built on the simple ability of organisations worldwide to independently publish content that uses URL-based hyperlinks to reference and attribute content published by others. Recent web standards now apply this directly to the *publication of data on the web*, by using extensible data schema, URLs for individual data items and open query APIs.

Localization is a Big Data industry that is poised to be transformed by open data on the web enabled by these new standards. In localisation, *words are our data*. Collections of terms and translations are commercially traded along value chains as part of localization projects. The leverage of terms and translations between projects is now being amplified by the uptake of data-driven language technologies applied to *machine translation and automated term extraction*.

Goals: The FALCON project combines the power of *open data* on the web with *data-driven language technologies* to construct the **Localization Web**. This consists of a network of terms and translations inter-linked to each other and to source and target documents via URLs. FALCON will integrate the resulting web of linked localisation and language data into localisation tool chains using existing data query and access control standards. Meta-data from these tools will add value to these data assets, enabling seamless *quality monitoring* across the value chain and their *on-demand leverage* in training machine translation and text analytics engines.

Partnership: FALCON does this by integrating the expertise of Ireland's CNGL Centre for Global Intelligent Content, with the technical and commercial knowhow of three international SMEs from the localization tool industry. From CNGL, *Trinity College Dublin* provides expertise in interoperable localization meta-data and *Dublin City University* provides expertise in machine translation (e.g. OpenMaTrEx.org) and text parsing. The commercial SME partners integrate

their existing Software-as-a-Service offering into a platform which is augmented with the linked data and language technology capabilities. *XTM International* integrates XTM Cloud, a web-based CAT tool and TMS that offers complete translation solution for project managers and translators. *Interverbum Technology* brings TermWeb, its multilingual, multimedia terminology management offering. *Easyling* integrates these with its award-winning web site translation platform.

Approach: FALCON will demonstrate the *active curation* of language resources and value-add meta-data, operating as an integral part of next generation localisation workflows. An *open meta-data schema* will capture the provenance of terms and translations as they progress through these workflows. The controlled, decentralized generation and sharing of this meta-data will yield new levels of *end-to-end visibility* into process and quality across the value chain. This will enable flexible, *on-demand assembly of training data* for targeted domain and *quality improvements* to machine translation and text analytics engines.

Benefits: As a result, clients and LSPs will be able to audit the use of language resources more precisely, providing a more reliable assessment of the return on investment in language technology. The use of open web data standards will provide opportunities for smaller players and public sectors bodies to leverage pooled resources while avoiding tool lock-in. Meanwhile, tool vendors and integrators will deliver offerings with expanded, future-proof language data integration features.

Get Involved! FALCON is actively seeking partners to engage in trialling of its technologies within commercial, public sector or crowd-source translation projects. Collaboration on standards and language technology integration are also sought. More details at: falcon-project.eu

FALCON is sponsored by the *European Commission* under grant 610879 from Oct'13 to Sept'15