Using Semantic Mappings to Manage Heterogeneity in XLIFF Interoperability

Dave Lewis, Rob Brennan, Alan Meehan, Declan O'Sullivan
CNGL Centre for Global Intelligent Content, Knowledge and Data Engineering Group,
School of Computer Science and Statistics, Trinity College Dublin, Ireland
dave.lewis@scss.tcd.ie, rob.brennan@scss.tcd.ie. meehanal@scss.tcd.ie, declan.osullivan@scss.tcd.ie

Abstract

The XLIFF 1.2 standard features several extension points that have served to complicate the full interoperability of translation content meta-data between tools from different vendors. Many vendors' specific extensions are in common use. XLIFF profiles promoted by individual large tool vendors or by consortia of smaller vendors (e.g. Interoperability Now!) attempt to reduce this complexity. However, as no one profile dominates, the overall result is that many XLIFF profiles are now in common use that extend the original standard in different ways. The XLIFF 2.0 standard attempts to control the evolution of extensions through the managed definition of new modules. However, until XLIFF 2.0 fully supplants the use of XLIFF 1.2 and its variants, tools vendors and language service providers will have to handle a range of different XLIFF formats and manage heterogeneity in the use of meta-data that impairs its use in automating steps in the localisation workflow.

Managing the mapping of different XLIFF profiles to an internal representation requires therefore, either extensive coding knowledge, or the understanding and maintenance of a wide range of different XSL Transforms. In this work we describe an alternative approach to handling the design, implementation and maintenance of meta-data mappings using semantic web technologies.

Keywords: *Semantic Mapping, Interoperability, Multilingual Web, XLIFF, RDF*

The localization industry is built on heterogeneous tool-chains with strong interoperability requirements. The XLIFF (XML Localization Interchange File Format) standard was established to enable greater interoperability between tools from different vendors. The XLIFF 1.2 (Savourel et al. 2008) standard has included several extension points to its structure with the aim to help provide greater interoperability between tools. However, these extensions have caused confusion among tool vendors and are rarely utilized. Instead, individual tool vendors have established their own extensions and as a result, many different extensions are in use causing complex interoperability issues. Specific XLIFF profiles, promoted by individual large tool vendors or by a consortium of smaller vendors, attempt to reduce the complexity and interoperability issues. Since no one profile dominates, the result is that many XLIFF profiles are now in use, which deviate from the XLIFF 1.2 standard in different ways. The XLIFF 2.0 (Comerford et al. 2014) standard attempts to control the evolution of existing extensions through the managed definition of new modules. Until the XLIFF 2.0 standard fully supplants the XLIFF 1.2 standard and the XLIFF profiles already in existence, tool vendors and language service providers still have to cope with the interoperability issues caused by the multitude of XLIFF formats in existence.

In this work, we present an alternative approach of overcoming XLIFF interoperability using Semantic Web technologies. In previous work (Lewis et al. 2012), a process is described how the use of Extensive Stylesheet Language Transformations (XSLT) (Kay 2007) at different points in the localization workflow can be used to uplift multilingual content and meta-data into a Resource Description Framework (RDF) (Cyganiak et al. 2014), Linked Data (Bizer et al. 2009) representation, also known as Linked Language and Localization data or L3Data for short. This provides a decentralized representation of the data, publishable on the web, where it can be shared among

localization enterprises for mutual benefit. Such benefits include access to a larger pool of language resources to aid in translation services and large datasets to train Statistical Machine Translation (SMT) tools. Interoperability issues are still present within the L3Data as multiple heterogeneous domain and tool-specific vocabularies are often employed within the RDF. However, the use of semantic mappings (Euzenat & Shvaiko 2013) can be employed to reduce this heterogeneity by transforming the L3Data from one vocabulary to another.

Our mapping representation, which we presented in (Meehan et al. 2014), is an RDF-based mapping representation that can be used to represent mappings between different L3Data vocabularies. The mapping representation uses a combination of SPARQL Inferencing Notation (SPIN) (Knublauch 2013) and meta-data. The executable specification associated with the mapping representation is a SPARQL (Harris & Seaborne 2013) construct query, which is executable on any standard SPARQL endpoint. The objective of the mapping representation is to provide a more agile approach to translation workflows and greater interoperability between software tools by allowing specific tool vendors to publish mappings, alongside the L3Data that they publish. This allows consumers of the L3Data to discover these mappings, through the use of SPARQL queries and execute them via a SPARQL processor.

Our use case is a Language Technology retraining workflow where publishing mappings leads to new opportunities for interoperability for the retraining of machine translation tools. Figure 1 below displays the process where a piece of HTML source content is acted upon by specific tools in a localization workflow. An XLIFF file is used to record the processing that the source content undergoes at each step of the workflow. At the end of the workflow, a custom tool using the XSLT language is used to uplift the data in the XLIFF file, to an L3Data representation, using the Global Intelligent Content (GLOBIC) semantic model (Brennan & Lewis 2014) vocabulary and store it in a triple store. This L3Data represents details such as the *source* and *target* of text content that underwent a Machine Translation (MT) process, which tool carried out the MT process, *post edits* and *quality estimates* associated with translated content. By building up L3Data in the triple store, it becomes a rich source of MT training data. The retraining aspect of the workflow involves retrieving content to be fed back into the SMT tool. This is achieved by querying the triple store for translated content with a quality estimate over a certain threshold value. SMT tools from different vendors, looking to utilize this L3Data for retraining purposes, need to have it mapped to a vocabulary they recognize. In Figure 1, the *MT tool* is unaware of the GLOBIC vocabulary, it is designed to consume data according to the Internationalization Tag Set (ITS) (Filip et al. 2013) vocabulary. The *Quality Estimate (QE)* and *Post Edited (PE) data* that is represented in GLOBIC must be mapped to an ITS representation for the *MT tool* to use it. Our mapping representation can be used in this situation since it is stored alongside the L3Data in the triple store. Mappings between the GLOBIC and ITS vocabularies can be discovered by a user/tool, through SPARQL queries

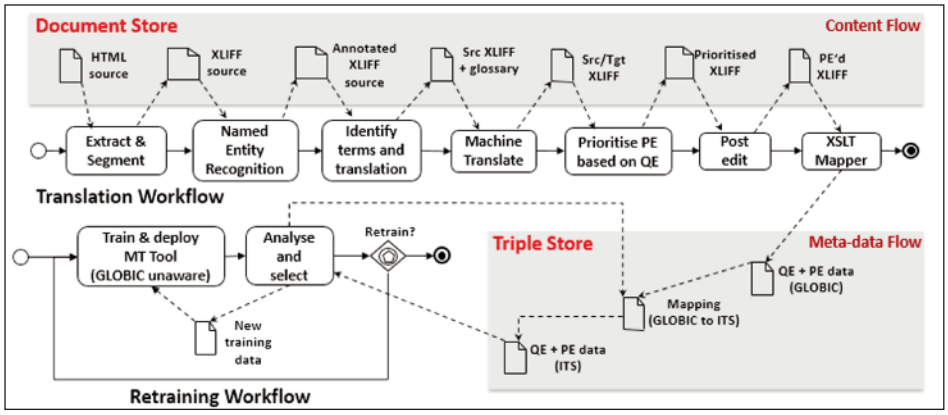

Figure 1. Language Technology Retraining Workflow

and executed. This will transform the L3Data, allowing the SMT tool to consume it.

References

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1-22.

Brennan, R. & Lewis, D. (2014). *The Global Intelligent Content Semantic Model Specification*. [Online] available: https://www.scss.tcd.ie/~meehanal/gic.ttl [Accessed 20 Jan 2015].

Comerford, T., Filip, D., Raya, R. & Savourei, Y. (2014). *XLIFF Version 2.0*. [Online] available: http://docs.oasis-open.org/xliff/xliff-core/v2.0/xliff-core-v2.0.html [Accessed 20 Jan 2015].

Cyganiak, R., Wood, D. & Lanthaler, M. (2014). *RDF 1.1 Concepts and Abstract Syntax*. [Online] available: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ [Accessed 20 Jan 2015].

Euzenat, J. & Shvaiko, P. (2013). Classifications of Ontology Matching Techniques. *Ontology Matching*, pp.79-84.

Filip, D. et al.(2013). *Internationalization Tag Set (ITS) Version 2.0*. [Online] available: http://www.w3.org/TR/its20/ [Accessed 20 Jan 2015].

Harris, S. & Seaborne, A. (2013). *SPARQL 1.1 Query Language*. [Online] available: http://www.w3.org/TR/sparql11-query/ [Accessed 20 Jan 2015].

Kay, M. (2007). *XSL Transformations (XSLT) Version 2.0*. [Online] available: http://www.w3.org/TR/xslt20/ [Accessed 20 Jan 2015].

Knublauch, H. (2013). *SPIN - SPARQL Syntax*. [Online] available: http://spinrdf.org/sp.html [Accessed 20 Jan 2015].

Lewis, D. et al. (2012). On Using Linked Data for Language Resource Sharing in the Long Tail of the Localisation Market., 2012. LREC.

Meehan, A., Brennan, R., Lewis, D. & O'Sullivan, D. (2014). Mapping Representation based on Metadata and SPIN for Localization Workflows. *In Proceedings of the Second International Workshop on Semantic Web Enterprise Adoption and Best Practice at ESWC*.

Savourel, Y., Reid, J., Jewtushenko, T. & Raya, R. (2008). *XLIFF Version 1.2*. [Online] available: http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html [Accessed 20 Jan 2015].