**1 Summary**

This implementation demonstrates how statistical machine translation (SMT) can automatically translate HTML documents from an ITS-conformant Web CMS.

In this use case ITS meta-data is use to solve the following problems:

* Informing the SMT service of precisely which sentences or sentence fragments should or should not be translated.
  + Benefit: Reduces the need for human checking to ensure the correct content has been translated using the correct language pair.
* Informing the SMT service, at a sentence or sentence fragment level, of the appropriate training corpora that would be appropriate for the SMT engine used.
  + Benefit: Improves the quality of machine translation by matching the training corpora of the SMT engine used as closely as possible to the type of text being translated.
* Proving the content manager with detailed provenance information on the outcomes of the SMT service invocation in terms of the engine used and the confidence score given to the
  + Benefit: Reduces the target language content quality assurance costs for the content manager, e.g. low scoring translations, or ones from an engine known to be less reliable, can be automatically extracted and passed for human translation review.
  + Benefit: Consistent quality problems with specific engines can be automatically correlated by the content manager for use in price/discount negotiations with the SMT service provider.

**2 Use Case Description**

This use case demonstration illustrates how ITS allows a HTML5 Content Author to communicate instructions on language, domain and translation to a simple segment-level statistical machine translation service provided by an MT Service Provider. It also shows how certain provenance information returned by such a service can be recorded in the target HTML5 document. This use case also highlights some issues in converting ITS mark-up into segment level content. The content application involves English reference content explaining terms and usage of a quote in another language (Latin), and also includes translatable quotes that may be better machine translated based on corpora of literature quotes.

This scenario may involve the following product classes: Content Authoring Tool; Source Quality Assurance (QA) Tool; Content Management System (CMS) and Web Browsers.

The business processes involved are: TBD

**3 Use Case Implementation**

The implementation of this use case involves the following components:

* [CMS-LION](http://www.w3.org/International/multilingualweb/lt/wiki/Simple_Segment_Machine_Translation_Use_Case_Demonstration#CMS-LION): This is developed by TCD under the CNGL project. It consists of an ITS parser and a simple segmenter that are integrated with a CMS (currently drupal). It is capable of performing round-trip interactions with translation tools via XLIFF or proprietary web services. Segment level changes recorded by each round-trip are recorded in an RDF based Provenance model. For this use case CMS-LION supports the following ITS2.0 data categories:
  + translate: HTML5, global and local
  + domain: HTML5, global
  + language information: HTML5, global
  + translationAgent: HTML5, global and local
  + mtConfidenceScore: HTML5, global and local
* [Segment-level, ITS-aware Matrex SMT Web Service](https://www.w3.org/International/multilingualweb/lt/wiki/Simple_Segment_Machine_Translation_Use_Case_Demonstration#Simple.2C_Segment-level.2C_ITS-Aware_SMT_Web_Service.3D): This is based on the Matrex SMT system developed at DCU as an extension to the MOSES SMT system. The web service interface is developed using a platform provided by the [PANACEA project](http://http:/www.panacea-lr.eu/). As a segment passed to this service is only a document fragment rather than a valid HMTL5 or XML document, this service is described as ITS-aware rather than ITS compliant. It uses proprietary web service parameters to relay the values of ITS data categories that apply to the whole segment, while they use a proprietary span element and attributes to relay sub-segment level ITS data categories. In this way the web service exhibits compatibility with the following ITS data categories: translate, domain, language information, translationAgent and mtConfidenceScore.

The operation of the system involves CMS-LION parsing and segmenting a HTML5 document containing ITS mark-up and then invoking the segment-level, ITS-aware Matrex SMT Web Service separately for each segment, then reconstructing a target language version of the HTML5 document, including relevant ITS mark-up. The interoperability points exposed in this use case are therefore: source HTML5 document, input to web service, output form web service and target HTML5 document

Limitations: This use case implementation provides an initial exploration of the integration of CMS and SMT highlighting segmentation related issues relate to SMT use, including the handling of differential ITS mark-up in sub-segments. Therefore, all ITS conformance behaviour is contained within the CMS-LION component. It is envisaged that a document level interface to Matrex will be developed in future, requiring ITS conformance by the service implementation.

**4 Use Case Demonstration**

* **Status**: Specification under development, implementation under development
* **Demonstration**:TBD.

**5 Interoperability Behaviour**

A step by step description of the demonstration, giving examples of how content and data is passed between components is visible at the interoperability points identified in the systems description. Any initial assumptions about the state of the system should be clearly stated. These examples should be consistent with each other from step to step so that the outcomes of various ITS-related processing can be clearly understood. A short explanation should be provided for each step, highlighting the role of the ITS data categories used.