Global Intelligent Content: Active Curation of Language Resources using Linked Data

David Lewis[1], Rob Brennan[1], Leroy Finn[1], Dominic Jones[1], Alan Meehan[1], Declan O'Sullivan[1], Felix Sasaki[2], Sebastian Hellman[3]

[1]CNGL at Trinity College Dublin, Ireland, [2]DFKI, Berlin, Germany, [3]University of Leipzig, Germany

E-mail: dave.lewis@cs.tcd.ie, rob.brennan@cs.tcd.ie, leroy.finn@scss.tcd.ie, dominic.jones@tcd.ie, meehanal@scss.tcd.ie, Declan.osullivan@scss.tcd.ie, felix.sasaki@dfki.de, hellmann@informatik.uni-leipzig.de

Abstract

As language resources start to become available in linked data formats, it becomes relevant to consider how linked data interoperability can play a role in active language processing workflows as well as for more static language resource publishing. This paper proposes that linked data may have a valuable role to play in tracking the use and generation of language resources in such workflows in order to assess and improve the performance of the language technologies that use the resources, based on feedback from the human involvement typically required within such processes. We refer to this as Active Curation of the language resources, since it is performed systematically over language processing workflows to continuously improve the quality of the resource in specific applications, rather than via dedicated curation steps. We use modern localisation workflows, i.e. assisted by machine translation and text analytics services, to explain how linked data can support such active curation. By referencing how a suitable linked data vocabulary can be assembled by combining existing linked data vocabularies and meta-data from other multilingual content processing annotations and tool exchange standards we aim to demonstrate the relative ease with which active curation can be deployed more broadly.

Keywords: Linked Data, Localization, Interoperability Standards

1. Introduction

Linked Data consists of fine grained, inter-linked data elements accessible via individual URLs (Bizer et al 2009). It provides an open data format that is shown to scale well to allow integration of data published by different organisation. This is already finding application in the publication of lexical semantic resources (Navigli and Ponzetto 2012). However, the more widespread use of linked data for gathering and exchanging language resources suitable for LT training has not yet been explored in detail. In this paper we present a linked data vocabulary designed to support the curation of training data for digital content that is richly annotated to support the requirements of modern internationalisation and localisation workflow. The ultimate aim is to generalise this to an approach for Global Intelligent Content that covers personalisation, information retrieval and multimodal interaction, but here just localisation is focussed. Global Intelligent Content is defined broadly as digital content augmented with knowledge that allows it to be aggregated and conveyed from creator to consumer. Workflows that process Global Intelligent Content are identified as a set of content processing components operating in different organisational contexts that collectively add value to Global Intelligent Content as it is conveyed from creator to consumer via some recordable content flow between components. At a systems level, the interoperability of Global Intelligent Content is manifested by interoperability between this involved interoperability between a number of Content Processing Components. A Content Processing Component is any system that adds value to Global Intelligent Content, or that allows other content processing components to improve the value they provide. It may consist of human elements, automated elements and/or other Content Processing Components in any combination. Where such Content Processing workflows combine language technology components and the application of human linguistic judgements, e.g. post-editing of machine translation, filtering of automated term extraction or human Wizard of Oz intervention in multimodal interactive dialogue systems, the semantic model supports the live, continuous curation of LT training data, termed Active Curation (Lewis et al 2012).

2. Active Curation in the Localization Industry

In this paper we specifically examine the application of this linked data vocabulary to the active curation of language resources within the content localization industry. At its most basic, this industry consists of content generating enterprises and the language service providers (LSPs) they contract to translate source content. In recent decades, the main technological innovations to yield productivity improvements have involved the collection and reuse of language resources. Specifically these take the form of: term-bases, which are multilingual glossaries that improve consistency in both authoring and translation of terms, and translation memories (TM), which are databases of previously translated sentences that assist translators in translating identical or similar sentences, phrases or terms. Clients able to provide a translation memory from a previous project will benefit from a translation project discount based on the closeness of the TM to the project text, using so called fuzzy match scores. Modern translation management system support live translation management and terminology databases, so that new segment and term translation from a translator or terminologist are immediately available to all others on the project. More recently, TMs and term-bases are being reused by LSPs as good quality training corpora for Statistical Machine Translation (SMT) engines, though the monetization of these benefits is less well established than TM fuzzy match discounts.

Localization can therefore already be regarded as a Big Data industry, where the monetized data exchange in the form of parallel text and term bases is well established commercially. However, the full benefits of commercial exchange for localization language resources is less available in the long tail of SMEs language service providers that make up the majority of localization market. Their low throughput of localized content means they have little opportunity to amass significant term-bases and TMs as assets to reuse between jobs or to train SMT engines tailored to their domain specialisms and language pairs. This is compounded by the lack of language resource curation skills needed to maintain these resources. The potential for the localization industry to benefit from pooling and exchanging language resources has already been recognized but only realized to date through centralized data sharing models, e.g. the repositories run by TAUS Data Association (www.tausdata.org) or LetsMT! (www.letsmt.eu). However, the sustainability of these approaches is limited either through the high cost of access, difficulty in predicting ROI or reliance on episodic public funding resulting in unpredictable levels of freshness, linguistic quality and data integrity.

To address this problem we develop a model to enable a decentralized approach to curating, locating and reusing language resources for the localization industry by using linked data. This approach allows language resource consumers to search and filter over distributed sources at different levels of granularity by using meta-data vocabularies and queries. This approach builds on the W3C's Semantic Web Standards; Resource Description Framework (RDF) and the Simple Protocol and RDF Query Language (SPARQL). These standards allow web content and web data (i.e. deep web content) to be interlinked in a decentralized and distributed manner, while remaining discoverable by sharing partners at any time through SPARQL queries.

We illustrate this approach with an example based on modern trends in language service provision. Firstly, language service provision is increasingly centred around web services and software as a service (SaaS) offerings. Online web site translation represents on growing trend with services offering translation and web site proxies in different languages by directly pulling content from the source language web site. In parallel, the tools used by translators, terminology managers and translation project managers are increasing available as SaaS offerings. Finally, we also see the growth in language technologies accessible as web services. These include machine translation and text analytics engines that may assist translators with tasks such as named entity recognition and disambiguation.

Figure 1 outlines how such service offering can be combined, with a client availing of the services of a multilingual web content management service. This service provider uses a SaaS translation management system to support its translators and project managers. This system uses a third party machine translation web service. It also integrates with a third party terminology management SaaS service, which in term uses a text analytics web service to assist in the correct identification and translation of terms in the text. This scenario avails of linked data is several ways.

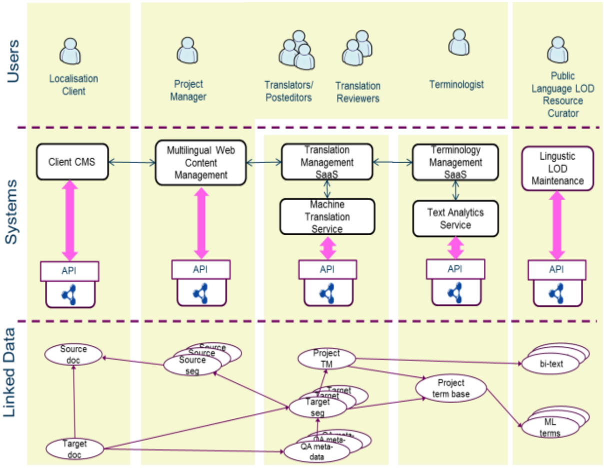


Figure 1: Example of link data in tracking localization workflows

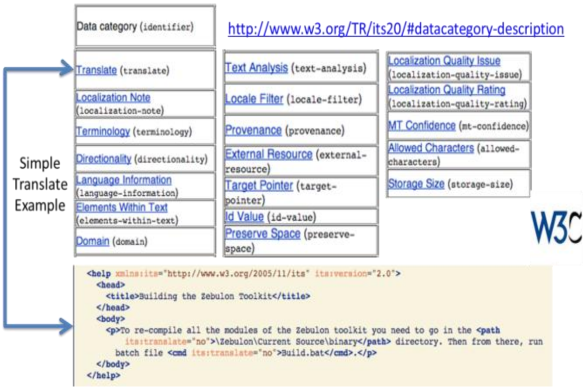
Firstly, lexical resources available as linked data already can provide a text analysis web service, e.g. Babel Net (Navigli et al 2013) and DBpedia Spotlight (Daiber et al 2013). These can return text annotation as references to nodes within their linked data graph, allowing the client application to search the graph more deeply if needed, e.g. to determine homonyms, morphologies and translations. Similar service exist for parallel text, e.g. the TAUS Data Association API for querying its large, shared translation memory. While to date, no such service operates from a linked data source, doing so may offer benefits in terms of being able to provide links to third party quality assessment of the parallel text on a per segment basis, and also for interlinking individual terms and phrases to lexical linked state resources, thereby enriching the resources for its clients.

This scenario are outlines however the further use of linked data for managing components within a workflow and for managing language resources generated by the workflow. Linked data can for example be used to record where the text analysis services were able to successfully identify terms in the source content and, by tracking through the manual translation step, where these provided accurate translation from linked lexical resources. Linked data can also maintain a persistent record of where reuse of the project translation memory was used in translating a segment or where machine translation was used and with what confidence. Where third party resources, such as lexical or parallel text resources were used, links can also be used to record such usage, therefore allowing project managers to ascertain the return of investment related to any costs in invoking these services. Where additional workflow tracking information can be used, such as the per segment time spent to translate or post-edit a segment, linked data may also simplify the task of correlating these human cost elements with the support provided by automated components such as machine translation and text analytics. This can further assist project managers in selecting and allocating both automated and human resources to particular projects, especially as performance may vary with the domain and style of the source content.

Finally, by building an interlinked trace of translation workflow execution and performance in this way, service providers will also systematically collect human linguistic judgments relevant to the language technology components. This can take to for of post-edits of machine translation or rejection of false-positives for text analytics components. This provides the basis for the process of active curation, whereby automated language technology components are systematically improved by continuous feedback of human corrections on a project by project basis. To build up such rich interlinked set so of linked linguistic data, such that they can be easily queried using the standard SPARQL query language, requires agreement between the developed of the individual components to record their usage according to a common vocabulary.

3. A Linked Data Vocabulary for Active Curation

Existing open vocabularies for recording process provenance and NLP corpora are used here to enable active curation and interlinking and reuse of languages resource across multiple linked data stores. We follow the decentralised evolution principles of open linked data, in making use of and extending existing RDF vocabularies, specifically making use of existing vocabularies for documents (e.g. Dublin core), the W3C Provenance Model (PROV) (Lebo et al 2013) and emerging linked data vocabularies for language resources, namely NLP Interchange Format (NIF) (hellman et al 2013). To extract content meta-data resulting from the various steps in the localisation workflow we rely on the rich set of multilingual content annotations recently standardised by the W3C Multilingual Web Language Technology working group in the Internationalization Tag Set (ITS) 2.0 specification. ITS2.0 addresses end-to-end meta-data issues in localisation workflow, with a specific focus on the role of language technologies such as machine translation, text analytics and automated quality assessment. The specification defines a set of abstract meta-data categories and defines how these can be implemented through a combination of attributes and elements in an HTML or XML document. Unlike existing XML and HTML meta-data attributes, ITS specifically annotates the textual content of document nodes rather than the nodes themselves. This allows the ITS annotation of the text to be reliably preserved as the content is transformed from one XML or HTML format to another.


Data category (identifier)
http://www.w3.org/TR/its20/#datacategory-description
Translate (translate)
Localization Note (localization-note)
Terminology (terminology)
Directionality (directionality)
Language Information (language-information)
Elements Within Text (elements-within-text)
Domain (domain)
Text Analysis (text-analysis)
Locale Filter (locale-filter)
Provenance (provenance)
External Resource (external-resource)
Target Pointer (target-pointer)
Id Value (id-value)
Preserve Space (preserve-space)
Localization Quality Issue (localization-quality-issue)
Localization Quality Rating (localization-quality-rating)
MT Confidence (mt-confidence)
Allowed Characters (allowed-characters)
Storage Size (storage-size)
Simple Translate Example
W3C

```
<help xmlns:its="http://www.w3.org/2005/11/its" its:version="2.0">
  <head>
    <title>Building the Zebulon Toolkit</title>
  </head>
  <body>
    <p>To re-compile all the modules of the Zebulon toolkit you need to go in the <path
       its:translate="no">\Zebulon\Current Source\binary</path> directory. Then from there, run
      batch file <cmd its:translate="no">Build.bat</cmd>.</p>
  </body>
</help>
```

Figure 2: ITS2.0 data categories and their use to annotate documents

Though it is not a normative part of the specification, ITS2.0 also defines a mapping from ITS conformant documents (in either XML or HTML) into RDF. Here, the ITS annotations are mapped into RDF properties with a subject representing the annotated text represented using the NIF vocabulary. A content processing provenance approach is taken to the core capability modelling of the Global Intelligent Content schema, where the PROV is used to express the state transformation that operates on content (principally on documents, terms, translation units and segments) and its metadata as the result of content processing activities, namely the humans, automation/tool and organisations involved, the activity type, and the activity timing. In the localisation workflow we examine, the Global Intelligent Content schema essentially consists of a combination of the PROV, ITS and NIF vocabularies.

We leverage the localisation industry's uptake of a standard bi-text interchange format, XLIFF, to provide the document format used throughout the workflow. A mapping of ITS onto XLIFF, combined with domain knowledge of typical XLIFF processing, allows us to automatically generate provenance logs annotated with ITS and NIF properties. It can also include activity specific attributes such as: confidence scores; quality assessments; fine-grained interaction (e.g. keystroke) logs from tools; license/copyright terms and links to other web resources, e.g. term bases or ontologies. The activity typing will support: authoring/revision; translatable content extraction and segmentation; source quality assurance (including terminology usage); Translation Memory leverage; SMT usage; human translation; post-editing and target language quality assurance. This approach can also be used to record value-adding operations to shared translation memory, term-base or parallel text elements, such as adding term translations, definitions or morphologies or identifying terms, or style and domain classification of entries. As this provenance-based approach serves to capture both the activity resulting in the recorded content transform or annotation and the agent responsible, it thereby supports the management of acknowledgement and credit for shared language resources, enabling auditing and return on investment calculation, as well as supporting sophisticated, license/copyright based access control through executable policy rules.

4. Conclusions and Further Work

This abstract outlines a linked data vocabulary that builds on existing standardised vocabularies to support provenance tracking of content and meta-data annotation resulting from localisation workflow. This supports both richer end-to-end workflow analytics but also offers a new approach to harvesting, enriching and exchanging language resources that has a ready commercial market across the existing language services industry. A prototype system is currently under development in the FALCON project (www.falcon-project.eu). This will serve to validate the current vocabulary and also to explore other specific extensions including, support for exposing generated language resources using an RDF version of the META-SHARE schema, providing project-level meta-data according to the Linport model (www.linport.org). Considering these other schema raises several important questions if we aim to interlink them in real world language services workflows.

One of the most pressing issues is the lack of a schema for describing language processing. The Linport schema describes the parameters of a localisation project, but defining the steps involved has been found problematic to accomplish due to the differing commercially competing views on the function of different activity steps. Similarly in the META-SHARE schema, a classification of usages that a language resource can be applied to has proved difficult to agree, in part because of the rapid rate of innovation in language content processing chains experienced in the language technology research community. The provenance based approach to linked open data may however provide a more direct approach to building language usage classifications. As it records what processes have been performed on a language resource, rather than those which may be performed in the future. This allows for a data-driven approach to defining service usages, based on the parameters of activities that have been previously executed. This allows the classification of usage to be built from comparison of activity instances rather than on theoretical discussions. Further, the need to capture sufficient activity provenance parameters as linked data also is encouraged by the ability to use this to record activity workflows in a repeatable manner (Garijo and Gil 2011).

5. Acknowledgements

This work has been partially supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) and partially by the European Commission as part of the MultilingualWeb-LT project (contract number 287815).

6. References

Bizer, Christian, Heath, Tom Berners-Lee, Tim (2009). "Linked Data—The Story So Far". International Journal on Semantic Web and Information Systems 5 (3): 1–22.

Joachim Daiber, Max Jakob, Chris Hokamp, Pablo N. Mendes (2013) Improving Efficiency and Accuracy in Multilingual Entity Extraction. Proceedings of the 9th International Conference on Semantic Systems (I-Semantics). Graz, Austria, 4–6 September 2013.

Daniel Garijo, Yolanda Gil, A new approach for publishing workflows: abstractions, standards, and linked data, Proceedings of the 6th ACM workshop on Workflows in support of large-scale science, Pages 47-56

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. , (2013) Integrating NLP using Linked Data. 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia

David Lewis, Alexander O'Connor, Andrzej Zydroń, Gerd Sjögren and Rahzeb Choudhury (2012) On Using Linked Data for Language Resource Sharing in the Long Tail of the Localisation Market, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)

R. Navigli and S. Ponzetto. (2012) BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250

R. Navigli, D. A. Jurgens, D. Vannella. (2013) SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. Proc. of 7th International Workshop on Semantic Evaluation (SemEval), in the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013), Atlanta, USA, June 14-15th, 2013, pp. 222-231

Timothy Lebo, Satya Sahoo, Deborah McGuinness, PROV-O: The PROV Ontology, W3C Recommendation 30 April 2013