The Role of Linked Data for Content Annotation and Translation

David Lewis, CNGL at Trinity College Dublin
Asunción Gómez-Pérez,
Sebastian Hellman, Universität Leipzig
Felix Sasaki, DFKI

Introduction: Content and metadata interoperability challenges

Increasingly, organizations, communities and individuals seek to access content not only in their own language, but also according to their own needs, preferences and context. Therefore content (such as news articles, product web sites, technical documentation) can no longer treated as monolithic and static artefacts. Fundamental challenges must be addressed, if content is to be dynamically created, curated, processed and delivered for consumers in global markets. The increased volume and velocity with which enterprises, institutions and users generate content requires new levels of automation to maximally leverage the limited capacity for language professionals to exercise appropriate linguistic judgments in processing content from creator to consumer, e.g. translating content or quality assuring content for consistency. Language technologies such as machines translation, text classification, and named entity recognition can support such automation, but only if used at the appropriate stages in the content processing chain and only if tailored to the characteristics of the content being processed and the need of the targeted consumers. A major interoperability challenge however is the variety in content formats used and in the annotation of lexical or semantic information that can be used to annotate and thereby ease the translation of words or phrases. This currently limits the efficiencies possible through language-technology automation, both in terms of consistently processing unstructured content and in training language technology to a particular content stream.

We identify the **Global Content Value Chain** as the business context for the processing of multilingual content from creation through to consumption (Emery et al, 2011). The central premise of the chain is that value can be added to content as it moves through the chain by leveraging of human judgments in combination with intelligent content service components. Today's Global Content Value Chain is best exemplified by the need to integrate between enterprise content management systems and the language services industry. Here workflows focus on enterprise-driven content creation, localization, management and publication functions. However, these value chains typically employ predefined workflows and complex decision making to pass content through the processing chain. The need to handle content variety often leads to specialization in the value chain, where companies -- often SMEs -- leverage either niche human skills (e.g. domain-specific translation in a certain language pair) or their specialized data (e.g. a specific domain lexicon or bi-lingual corpora). This specialization however heightens the need for smooth **interoperability** since otherwise the overhead of manual intervention required for the exchange and processing of content will inhibit the growth of

the market.

When considering content communicated via the web, it will typically consist of **unstructured content** such as text, audio or video **accompanied by structuring markup and by metadata** which serves to annotate both unstructured content and the markup. The markup and metadata annotations play a key role in the processing of content, including its transport, indexing, aggregation, selection, filtering, adaptation, composition and presentation. Content interoperability therefore relies on a common understanding of how to process the content markup and metadata annotations that can be shared between different content processing components. It is therefore the extant variety of content markup and annotation techniques that makes content interoperability complicated and often expensive to achieve.

The importance of **data provenance and provenance about the processes that generate, consume or use data** have increased dramatically in the last decade, due to the increasing availability and demand for metadata in a wide range of applications in many fields ranging from social networking, scientific workflows, machine translation workflows, etc. Being able to track how a piece of information has achieved its current state is a great help for any user to understand where that information comes from, how it has been generated, or how a certain result from an experiment can be rebuilt. However metadata per se is useless; we need to know how it affects the data and how it relates to the different pieces of information and processes to understand their origin.

Annotation Patterns for Multilingual Web Content

For Web Content, interoperability has been considerably eased by the widespread adoption of document formats that adopts tree-based serializations. This has enabled a common programmatic abstraction for document processing to be standardized in the form of the DOM, Document Object Model (Le Hors et al 2004). This in turn has enabled development of common declarative mechanisms for selecting tree nodes within a document (Clarke & deRose 1999) and performing transformations on document contents (Clarke 1999). This again has in turn proved powerful in developing content processing chains in enterprise content applications, which typically span web, print and other content delivery channels. However, for native web content applications these benefits have been diluted somewhat in the drive towards HTML5, which has integrated several elements that dilute common DOM serialization of content to bring benefits of enhanced interactivity and rich content media delivery, e.g. ECMA Script, audio and video content formats.

In addition, the Web has experienced the growth of the semantic web and interest in its potential role in content discovery and delivery. The semantic web offers a fine grained graph of data nodes accessible as web resources, i.e. by dereferencing a URI, together with navigable links between these data resources. This has enabled newly standardized mechanisms, such as RDFa (Herman 2013) and schema.org, to be employed for interlinking linking web resources in the form of content-bearing documents and external metadata in the form of linked data nodes. The result is a rich but complex set of mechanisms that can be employed in content processing

and which therefore must be accommodated when attempting to implement efficient integration of content processing components into content processing value chains.

The translation and localisation of content is key to the generation of multilingual content on the web and possesses it own specific interoperability problems. The various stages of the translation process, e.g. machine translation, reuse from translation databases, post-editing to correct automated translation, human translation and translation quality review, may be undertaken by different workers, with service providers using different tools and automated processing components. It is therefore important that content and its translations are passed reliably between such bi-lingual content processing tools. OASIS has developed the XML Localization Interchange File Format (XLIFF) standard to address this (Savourel et al 2008). XLIFF offers a bi-text exchange format that accommodates a wide range of metadata needed for the localization process, including integration of translation reuse, human post-editing of automated translation and terminology management.

However, annotation related the processing of multilingual content often need to be performed on the source authoring and target language publishing format. The W3C has addressed this by defining standard metadata for annotating content as independent data categories that annotate existing DOM-based formats either for stand-alone use cases, or used in combination to support interoperability across the content processing chain, regardless of mapping between different formats used within it. This was initially defined in the Internationalization Tag Set specification (ITS1.0) (Lieske & Sasaki 2007), which has recently be supplemented by ITS2.0 (Filip et al 2013). This expands the implementations of ITS from just XML to include HTML5 and RDF. The key insight, continued from ITS 1.0, is that the data being annotated is the textual content of documents. This enables ITS to be used with source and target language documents or with XLIFF as a XML bi-text exchange format. A specific profile for using ITS2.0 with XLIFF, clarifying how to represent ITS2.0 data categories within XLIFF, has been defined[1].
Common annotation schemes oriented toward the semantic web and linked open data, i.e. RDFa and microdata, are not well suited to this task as text is treated only as literal objects of data triples and not the subject of metadata annotations. Hence, the RDF representation of ITS 2.0 relies on NIF (Hellmann et al. 2013): The NLP Interchange Format is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. NIF allows to create URIs that identify ranges within text string. In this way NIF provides important features for which the aforehand mentioned annotation schemes are not well suited.

By generalizing over the ITS 2.0 content annotation approach, the following patterns can be identified.

Pattern	**Description and advantages**

[1] http://www.w3.org/International/its/wiki/XLIFF_1.2_Mapping

P1. Annotation of Textual Content in a DOM conformant document	The annotation of text nodes (i.e. the textual content of element nodes) and the textual content of attribute nodes in a DOM conformant document can be specified by association with well-defined attribute nodes. This is the base pattern of this pattern language and has the advantage of using well defined attributes to specify annotations allows these annotations to co-exist with other DOM conformant schemas in a variety of applications
P2 Direct sub-tree annotation	An advantage of this approach is that this element can be placed outside of the main content-bearing portion of a document, e.g. in the <head> element of a HTML document. This approach also offers the flexibility to easily annotate a non-contiguous set of parts of a document.
P3 Selector-based annotation	This pattern exploits the standardized specification of node selector language that can operate with DOM-conformant language, such as XPath and CSS selectors. An annotation therefore can be associated with a set of nodes by associating it with a selector statement that specifically identifies that set of nodes.
P4 Referenced External Selector-based Annotation	Selector based annotation rule can be defined in an external file that can be referenced from within a document that uses those rules for annotation. This has the advantage that the same set of rules can be easily applied in a consistent manner to a whole set of the document. This is useful, for example, when the rules define annotations that relate to a schema used by a number of documents. It also allows the rule in the references files to be modified without altering the referencing files.
P5 External binding to selector-based annotation	An external definition to selector based annotation may also be bound externally to a document, e.g. by using metadata in a CMS content repository.The advantage of this is that the binding can occur with no impact on the structure and content of the document.
P6 Referenced Annotation	Here the annotation is not held in an attribute value, but instead the attribute specifies an Internationalized Resource Identified (IRI) that can be dereferenced (typically retrieved with a HTTP GET) to yield the metadata value. The advantage of this pattern is that the IRI can point to structured data so that annotation of a more complex type than is permitted in attribute node values can be used. The value of the annotation could in fact be any media or media fragment type, from a fragment in a DOM-conformant document, to an RDF node or even rich media content such as an audio or video resource.
P7 Pointer Pattern	Metadata that can be used to annotate text nodes may sometimes already exist in the document, but as an ad hoc text node or attribute node value, which is therefore difficult to parse in an interoperable way. This pattern makes explicit that another part of the document can be used to annotate textual content. The advantage of this pattern is that it allows existing piece metadata to be reused to provide interoperable textual annotation with a minimal impact on the document
P8 Multi-Annotated Text	The lack of ordering semantic for attributes in a DOM conformant document means that only one attribute node of a given name may be associated with a given element node. However in some circumstances an annotation of a given type may need to be applied several times to some text in a document. This may be because we wish to record that different values for an

	annotation where applied at different points in time, or that different annotating agents had different views on what the value of the annotation should be. This can be done using nested enclosing elements, such as html spans, or using values attributes with multivalued text encoding or capture multiple values in subelements of a separate multivalues element, which is reference by the element attribute of the annotated text.
P9 Annotation metadata	This pattern allows the annotation itself to be associated with additional metadata. This is useful if the way in which the annotation was generated has a bearing on how it should be interpreted. ITS 2.0 uses this pattern to associate a reference to the engine that has generated an annotation containing a confidence score with that annotation's data category. This is important since confidence scores are not comparable across engines, so identifying the engine involved is key to making use of the score
P10 External annotation of document fragments	A document may also be annotated by externally associating external metadata with a fragment identifier in the document.

In relation to linked data, several ITS 2.0 data categories make use of pattern P6 to link to external meta data via a URL reference, which could point to a linguistic linked data resource. In particular the Text Analysis and Terminology data categories can point to lexical resources, e.g. a node in BabelNet.org. Also of interest is how records of the language technology engine used (P9) can be established (which is supported by the annotatorsRef feature in ITS2.0) and also how application of language technology, e.g. text analysis and machine translation can reveal more information about who, what and when the processing occurred. We will examine this in more detail below.

Provenance for Multilingual Web Content

Due to the different nature of the algorithms, tools and data involved in the machine translation activities, **provenance constitutes an important source of information about the execution of processes and machine translation workflows**. In fact, provenance can support reproducibility of translations, experiments and a better interpretation of the different stages of the whole process and data sources used. However, the provenance about every activity can be provided by different systems, by using different models and granularity.

In order to properly capture provenance, models and vocabularies with different granularity have been developed. Some are coarse-grained and focus on capturing the process in which an object has achieved its current state, as in workflows or scientific experiments (e.g. the Open Provenance Model (Moreau et al 2011), Provenir (Satya et al 2010), or PML (Da Silva et al 2006) (McGuinness et al 2007); while others are more fine grained, capturing the license, signature or attribution of a document (e.g. Dublin Core (Dublin 2009), Premis (Higgins 2008) or SWAN[i]). There are also vocabularies focused on suggestions about how to publish the provenance information of LD (e.g. Provenance Vocabulary (Hartig 2009)). Finally, the W3C Provenance Working Group[ii] is concerned about these current issues, and they have defined

the PROV standard[iii], a family of specifications that aim to provide the means to describe the provenance of a resource independently of the context in which they are defined.

The integration of provenance across heterogeneous systems is a challenge in machine translation, requiring both conceptual breakthroughs and standardization efforts. For instance, within this topic an important challenge would be the definition of queries over provenance obtained from heterogeneous sources and processes. Formal models of provenance in different settings, and unifying formalisms for cross-layer provenance, are needed to make it easier to querying heterogeneous provenance arising from different data or computational models.

References.

Clarke, J., (1999) XSL Transformations (XSLT) Version 1.0, W3C Recommendation 16 November 1999

Clark, J, DeRose, S., (1999) XML Path Language (XPath), Version 1.0, W3C Recommendation 16 November 1999

Dublin core metadata initiative. DLib Magazine, 6(12), 2009.

Emery, V., Kadie, K., Laplante, M. (2011) Multilingual Marketing Content: Growing International Business with Global Content Value Chains, Content Globalization Practice Research Report, Outsell, 2011

Filip, D., McCance, S., Lewis, D., Lieske, C., Lommel, A., Kosek, J., Sasaki, F., Savourel, Y., (2013) Internationalization Tag Set (ITS) Version 2.0, W3C Proposed Recommendation 24 September 2013, accessed from http://www.w3.org/TR/its20/ 29th Oct'13

Olaf Hartig. Provenance Information in the Web of Data. In Proceedings of the Linked Data on the Web (LDOW) Workshop at WWW, Madrid, Spain, 2009.

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

Herman, I., et al RDFa 1.1 Primer - Second Edition, Rich Structured Data Markup for Web Documents, W3C Working Group Note 22 August 2013

Sarah Higgins. Premis data dictionary for preservation metadata. (March):224, 2008.

Savourel, Y., Reid, J., Jewtushenko, T., Raya, R.M., (2008), XLIFF Version 1.2. OASIS Standard 1 February 2008, access from http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html 29th Oct'13

Le Hors, A., et al (2004) Document Object Model (DOM) Level 3 Core Specification, Version 1.0, W3C Recommendation 07 April 2004

Lieske, C., Sasaki, F., 2007, Internationalization Tag Set (ITS) Version 1.0, W3C

Recommendation 03 April 2007, accessed from http://www.w3.org/TR/its/ 29th Oct'13

Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul T. Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric G. Stephan, and Jan Van den Bussche In: Future Generation Comp. Syst., Vol. 27, Nr. 6 (2011) , p. 743-756.

Deborah L McGuinness, Li Ding, Paulo Pinheiro Da Silva, and Cynthia Chang. PML 2: A Modular Explanation Interlingua. In Proceedings of the AAAI'07 Workshop on Explanation-Aware Computing. Knowledge Systems Laboratory, Stanford University, 2007.

Paulo Pinheiro Da Silva, Deborah L McGuinness, and Richard Fikes. A Proof Markup Language for Semantic Web Services. Information Systems, 31(4-5):381–395, 2006

Satya S. Sahoo, Roger Barga, Amith Sheth, Krishnaprasad Thirunarayan, and Pascal Hitzler. PrOM : A Semantic Web Framework for Provenance Management in Science. 2010.

.

---

[i] http://swan.mindinformatics.org/v2.html
[ii] http://www.w3.org/2011/prov/wiki/Main_Page
[iii] http://www.w3.org/TR/2013/REC-prov-dm-20130430/

CVs

David Lewis is Director of Knowledge and Data Engineering Group (KDEG) in TCD. He is a principle investigator of the CNGL Centre for Global Intelligent Content (www.cngl.ie), leading research into the interoperability and analysis of multilingual, multimedia and multimodal process chains. He coordinated the FALCON project (falcon-project.eu) which is prototyping the integration of linguistic linked data into localisation tool chains, and he also participates in the LIDER project. He is currently co-chair of the MultiLingualWeb - Language Technology Working Group at the W3C and on the advisory board of GALA's Collaborative Research, Innovation and Standards Program (CRISP).

Asunción Gómez-Pérez is Full Professor at UPM, director of the Artificial Intelligence department, director of the OEG and PhD in Computer Science (1993). Before joining UPM, she was visiting (1994-1995) the Knowledge Systems Laboratory at Stanford University. She also was the Executive Director (1995-1998) of the AI Laboratory at the School of Computer Science.

She coordinates LIDER and she has coordinated SEALS, SemSorGrid4Env and Ontogrid and she has participated in more than 15 EU projects. Her main research interests are ontologies, Linked Data and the Semantic Web.

Sebastian Hellmann (AKSW, Universität Leipzig, Germany, hellmann@informatik.uni-leipzig.de, http://bis.informatik.uni-leipzig.de/SebastianHellmann) finished his Master thesis in 2008 at University of Leipzig and is currently a research fellow for the AKSW research group and also a member of the LOD2 EU project. He is contributor, co-founder and leader of several open source projects including DL-Learner, DBpedia, and NLP2RDF. Sebastian is author of over 15 peer-reviewed scientific publications and was chair at the Open Knowledge Conference in 2011, the Workshop on Linked Data in Linguistics 2012, the Linked Data Cup 2012 and the Multilingual Linked Data for Enterprises 2012 workshop.

Felix Sasaki works within the W3C Internationalization Activity, which assures that the needs of Internationalization are fulfilled within W3C-specifications, and also creates Internationalization specific specifications. Felix is the team contact for the W3C Internationalization Core Working Group and the ITS Working Group. He is used to crossing borders between cultures and (scientific) communities: He studied Japanese, Linguistics and Web technologies at various Universities in Germany and Japan. Now his mission is to address the needs of Internationalization and Localization within the W3C, applying different W3C technologies like RDF and XML for purposes of Internationalization and Localization, and introducing the merit of these technologies to a broad Internationalization and Localization community.