Generalizing ITS as an Interoperable Annotation Technique for Global Intelligent Content

Dave Lewis, Leroy Finn, Rob Brennan, Declan O'Sullivan and Alex O'Connor
Centre for Next Generation Localisation
Trinity College Dublin, Ireland
dave.lewis@scss.tcd.ie, *finle@tcd.ie*, *rob.brennan@scss.tcd.ie*, *Declan.osullivan@scss.tcd.ie*
alex.oconnor@scss.tcd.ie

Abstract

This paper considers how the interoperable content annotation techniques developed to address the needs of localization processing chains could be applied to a broader class of content processing. We extract the content annotation patterns developed for the Internationalization Tag Set standards at the W3C. These provide a means for annotating content with common meta-data that addresses different aspects of content localization from content creation, through extraction, segmentation, terminology management, automated translation, post-editing, quality assurance to publication of the translated content. This paper explores the lessons learnt in developing ITS2.0 as a suite of interoperable content annotation in the form of a pattern language. Interoperability problems arise when end-to-end content processing spans different: content formats; content processing tools and engines; and content processing service providers. This paper aims to make it easier to leverage these annotation patterns in the same way across these different interoperability mechanisms. In particular we propose annotations that follow the ITS annotation patterns but address personalization content processing. From this proposal, the potential for integrated localization and personalization processing is considered.

1. Introduction

One of the most significant changes to people's lives in recent years has been the explosion of content available to users, enterprises and communities via the Web. Enterprises and users have adopted new roles as creators, curators and consumers of content, in social and corporate contexts. 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. Fundamental challenges must be addressed, however, if content is to be dynamically created, curated, processed and delivered for consumers in global markets. The content processing value chains that deliver content from creators to consumer must address the volume, velocity and variety of content. 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 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 that exists in content formats used and in the linguistic domains, lexica and styles exhibited by content. 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 propose new unifying concept called 'Global Intelligent Content' as a basis for addressing these interoperability challenges. This concepts calls for embedding new levels of interoperable knowledge and intelligence into content to enable advanced intelligent content services to automatically process and transform that content in a more consistent and responsive manner. These intelligent content services will combine data driven language technologies and semantic reasoning capabilities. In this way, Global Intelligent Content will be more discoverable, semantically rich, adaptable, contextually aware and reusable across different granularities across global markets, right down to the individual. Global Intelligent Content should therefore be dynamically transformed based on current user interaction, perceived user intention or current delivery context.

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 niche human skills (e.g. domain-specific translation in a certain language pair) or the specialized knowledge needed to leverage specific language resources using language technologies, 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.

In this paper we examine the interoperability requirements of two important classes of content processing that we regard as key to the formation of global content management chains, namely *localization and personalization*. Localization is the industrial process of adapting content to a target locale. This is primarily concerned with the translation of textual content, but may also involve the adaptation of images; currency, date and other data formats and layouts to the norms of the target market. Personalization describes a range of techniques used to adapt content to an individual user's needs. It depends on a user model and employs techniques of navigation adaptation (hiding or prioritization of hyperlinks), adaptive discovery (adapting content indexing and queries), content adaptation (e.g. selection and filtering of content elements) and content composition (Levacher at al 2009, Koidl et al 2011, Wade 2009). Currently, Localization is the more mature field in terms of interoperability standards. We therefore review existing approach to standards to examine the content annotation solutions they offer that might best provide common content meta-data that may persist across a workflow of heterogeneous components. From this analysis we see that the approach to content annotation defined in the Internationalization Tag Set (ITS) standard from the W3C (Savourel et al 2008) best addresses the needs of interoperable content annotation. We then extract these annotation techniques, based on the current ITS2.0 specification, to generate a set of reuable annotation patterns. We end by proposing new personalization-specific meta-data that could exploit these patterns to provide interoperable content meta-data annotation specifications.

2. Content Interoperability Challenges

At its simplest, content can be regarded as digital media specifically created by people with the express intent to be consumed by other people (thereby allowing us to distinguish it from digital data either solely generated or solely consumed by automated systems). 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 meta-data which serves to annotate both unstructured content and the markup. The mark-up and annotating meta-data plays 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 annotation that can be shared between different content processing components. It is therefore the extant variety of content mark-up and annotation techniques that makes content interoperability complicated and often expensive to achieve when attempting to form real world content processing chains.

If we consider content on the Web in particular, 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 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 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 format.

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[1], to be employed for interlinking linking web resources in the form of content-bearing documents and external meta-data 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.

This is particularly challenging to the classes of content processing that we are considering in this paper, namely localization and personalization. Both often suffer in practice from being employed in a post-hoc manner, such that downstream localization and personalization processing is not adequately considered in the up-stream content processes where content is created, structured and annotated. This therefore adds to the cost and complexity of localization and personalization processes as they must accommodate and often also preserve the diversity of content mark-up and annotation as they traverse these downstream processes. This is required in order to maintain the validity of assumptions about mark-up and meta-data made in subsequent downstream processing components involved in content publication, indexing, search engine optimization, archiving and reuse. Therefore, making extensive changes to the mark-up of content to accommodate localization or personalization processing may not be attractive option for enterprise. In the first instance this is because it would prove too disruptive to other downstream processes (including between personalization and localization processes). Also, such changes may result in personalized and/or localized content being ‘forked’ away from parallel versions of the same content passing through pre-existing content processing chain (e.g. for print publication or search indexing), making it difficult to recombine or reuse that content in future iterations. For this reason, we therefore focus here on the mechanisms available for annotating content for localization and personalization, rather than consider alternative mark-up formats that would ultimately be more difficult to deploy in the context of existing content value chains.

The next section examines the start of the art in open, interoperable content mark-up and annotation specifications for the more mature field of localization, in terms of their capabilities for marking up and annotating content.

3. Analysis of Content Interoperability Mechanisms for Localization

Localization is a well-established part of the content processing chain for many multinational companies. However, content processing value chains involving localization workflows can be varied and complex and overheads due to poor data and meta-data interoperability are estimated as being upto 20%. Moreover, the distribution of providers by size exhibits an extremely long tail, with 99% being SMEs, who therefore struggle to both handle the overhead of poor interoperability and to reap the benefits of large scale language data reuse arising from large volumes of translation traffic.

The localization 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 in this industry have involved the collection and reuse of language data resources. Specifically these resources take the form of: term-bases (multilingual glossaries that improve consistency in both authoring and translation of terms) and translation memories (databases of previously translated sentences that assist translators in translating identical or similar sentences, phrases or terms). The leverage of translation memories is supported by a well-established norms for translation discounts based on the corresponding human translation effort savings. More recently, translation memories (TM) and term-bases are being reused by LSPs as good quality training corpora for Statistical Machine Translation (SMT) engines. Therefore the

[1] http://schema.org/docs/gs.html

collection, distribution and reuse of both parallel text and bi-lingual term bases is a key part of the localization workflow.

Poor interoperability experience arise in many localization workflows due to the multiple parties involved using a variety of content formats, workflow systems and translation tools. Though there are several standards serving this industry, standardization efforts are somewhat fragmented between several different organizations.

To avoid this fragmentation disrupting our analysis, the interoperability standards examined below are categorized by the type of interoperability function they perform.

Content authoring and publication formats: These include standardized electronic publication formats such as HTML (Berjon et al 2013), OASIS DITA (Eberlein et al 2010) and DocBook[2]. There is however widespread usage of content authoring and publication formats are open in that the specification is published, but are proprietary in that the design of the format is not subject to a consensus forming process that is open to broad industry input and consultation. Examples are PDF, Rich Text Format, Microsoft Office and Open Office formats and Adobe XX formats. Often, authoring is performed in a different format to publication, where HTML and PDF have become dominant. This requirement has made XML content authoring formats more popular, as XSLT declarations can be used and exchanged to offer reliable transforms for authoring to one or more publication formats. This in turn promotes the uptake of component or topic based authoring, where content is authored in discrete units designed to be easily recombined at the publication stage. These formats are not primarily focused on the needs of localization, sometime then requiring supplementary annotations for internationalization and localization purposes. This has been somewhat addressed by the W3C through the standardization of the Internationalization Tag Set (ITS v1.0). This aims to reduce elements of the interoperability overhead cost by defining a set of well-defined independent standard meta-data attributes that can be used to annotate XML content to address specific use cases. These use cases are: whether to translate content or not; where content is a term or not; identifying subflow in text to assist translators; offering localization notes for the translator; providing language information when absent in the source format; and providing directionality and ruby annotation information often needed in non-latin scripts. So while the wide range of source content format is a major source of complexity in localization content processing chains, as ITS is agnostic of the XML format used for the source it can be used consistently, in concert with conformant ITS processors, across any XML format, including bi-text exchange formats discussed below. Further, it defines its annotation, known as data categories, in an abstract manner that is independent of the XML implementation and could be potentially applied to other non-XML formats, though this is not yet in common practice. Addressing this requires the development of content extraction filters, which are needed because the translation processes is performed largely separately from the content authoring and publication processes. This makes translating content in the context of the publication format problematic and also complicates the synchronization of translation processes with ongoing changes made to the source content. The development and maintenance of extraction filters is a complex task, with limited support for open solutions, meaning that extraction components must be developed and used in tandem with reassembly components. Defining content annotation that can be easily process in content filters is therefore an important objective of ITS.

Language resources: The reuse and leverage of language resources is a key productivity driver in localization processes. Principle amongst these is translation memory, which provides a searchable database of previous translations to avoid effort in replicating similar translation. The Translation Memory Exchange (TMX) standard provides an XML vocabulary for exchanging parallel text (or bi-text) that capture source language content and its translation at the level of segments as used in the translation processes that generated them. TMX is well supported in translation management systems (TMS) and computer assisted translation (CAT) tools. The widespread use of TMX has also prompted its increasing use as a format of providing parallel text to processes training statistical machine translation components. Consistent use of terminology from authoring to translation (human and machine based) and translation review is important in achieving good quality translation. Within the localization process exchange of this information between tools in the form of term bases is supported by the Term Base eXchange XML vocabulary (TBX). ISO has been active in promoting open formats for lexical repositories. In recent years, mapping of these lexical repository formats into the Resource Description Framework (Manola & Miller 2004) that underlies the semantic

[2] http://www.docbook.org/

web, for publishing as linked open data have been explored (Windhouwer & Wright 2012). Other, RDF vocabularies have been proposed for publishing of lexical resources directly as linked open data (Chiarcos 2008, Buitelaar et al 2008). In parallel large open cross lingual and lexical repositories are emerging, based on existing resources such as Wikipedia and WordNet, with their increasing usage presenting de facto standardization of their vocabularies – reflecting an increasing trend in the development of common formats in the linked open data community.

As natural language technologies have become increasingly viable, there has also been interest in developing language resource formats that can convey the output of language processing, including lexical parsing, semantic tagging and named entity recognition. This has resulted in a proposal for an RDF vocabulary supporting the exporting of language resource resulting from NLP component processing, termed the NLP Interchange Format (NIF) (Hellman et al 2013).

Bi-lingual Tool exchange formats: The various stages of the translation process, e.g. machine translation, TM leverage, post-editing, human translation and translation review, may be undertaken by different workers, service provider each using different tools and processing components. It is therefore important that content and its translations to be passed between reliably between such bi-lingual content processing tools. One approach popular in software UI translation is the user of the PO format for passing translatable content to translation processes and be returned matched with translation. Though a popular format, especially in open source software projects, it does not benefit from an open industry agreement process. A more concerted standardization has been conducted by OASIS in the development of XML Localization Interchange File Format (XLIFF) (Savourel et al 2008). This offers a bi-text exchange format that accommodates a wide range of meta-data needed for the localization process, including integration of TM leverage, human post-editing, translation and review and terminology.

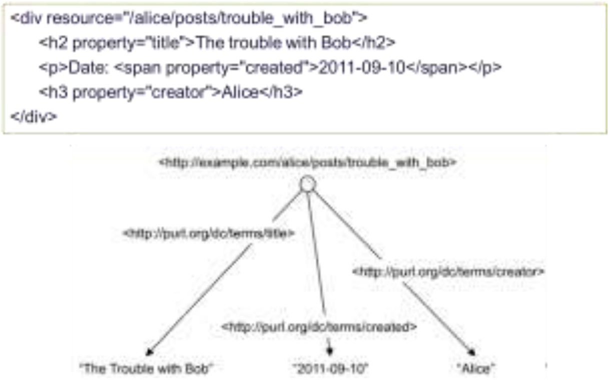
Processing instructions: The effectiveness and fidelity of a localization process chain is particularly sensitive to how certain processes are conducted. In such cases having the ability to exchange instructions between tools and worker in an open format is important. One of the most crucial process instructions is the segmentation of text into translatable segment, since efficient leverage of translation memories requires consistent segmentation. The Segmentation Rule Exchange (SRX) format allows such rules to be exchanged and segmentation outcomes to therefore be accurately reproduced between tools.

Discussion

It can be seen from this brief analysis that interoperability formats for localization suffer from fragmentation in goals, the bodies that produce them, the formats they use, the use case they address and their uptake within the localization process. Two recent initiative has attempted to address this fragmentation.

A small industrial consortium, known as ‘Interoperability Now!’ (IN!), has formed specifically for the task of developing a Translation Interchange File Format. This defines how several related open formats can be packages and zipped for exchange between tools, including XLIFF, TMX and TBX. While this performs a useful consolidation function, it has progressing in parallel with a revision of the XLIFF standard with many of the same goals, including the restriction of options that was perceived to slow uptake of XLIFF 1.2. In this sense IN! has also served to add to the sense of fragmentation in the industry. A key factor here, which is similar phenomenon in web services interoperability, is that the ease with which a format can be extended using name spaces means that the key concepts represented by the format can be changed. This adds unforeseen complexities to the updates required to third party components intending to implement the extension. The key here is to ensure that the semantic role of different format elements is clearly defined separately from the syntax of the format – however this is a complex task to achieve in practice. The result is a complex set of interlinked XML vocabularies that are carefully tuned to the need of localization process interoperability, but which as a result is poorly suited to more general content processing.

The other initiative has been the Multilingual Web – Language Technology at the W3C. Rather than attempting to develop a broader container format, it takes to approach adopted in ITS1.0 to define independent data categories that annotate existing 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. The result is a draft ITS2.0 specification (Filip et al 2013). This expands the implementations of ITS from just XML to include HTML5 and RDF. The key insight, continued from ITS1.0, is that the data being annotated is the textual content of documents. 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 meta-data annotations as outlined in figure 1.



```
<div resource="/alice/posts/trouble_with_bob">
    <h2 property="title">The trouble with Bob</h2>
    <p>Date: <span property="created">2011-09-10</span></p>
    <h3 property="creator">Alice</h3>
</div>
```

<http://example.com/alice/posts/trouble_with_bob>
<http://purl.org/dc/terms/title>
<http://purl.org/dc/terms/creator>
<http://purl.org/dc/terms/created>
"The Trouble with Bob"
"2011-09-10"
"Alice"

Figure 1: RDFa content annotation

However, to support close integration with content processing and localization tool chains, ITS associated meta-data with textual content either through well-defined attribute added to enclosing elements (e.g. HTML span) or through rule element that associate attributes with enclosing elements (or attributes) using XPath selectors. Well defined inheritance, override and default rules enable dedicated ITS processor functions to be implemented and conformance tests for such processors to be formulated. Ease of adoption is supported by conformance being attainable through implementation of a single data category, presenting a lower cost migration path than the wholesale adoption of a specific source or bi-text interchange format. In addition to the data categories in ITS1.0, ITS2.0 adds further data categories designed to ease the integration of language technologies and linked open data into the localization process. Machine translation integration is supported by annotation of the content's application domain and of automated translation confidence scores. Text analysis is supported with annotation to associate words or phrases with external resources, e.g. DBpedia for classification and definitions or WordNet or BabelNet for lexical definitions. Such annotation may be generated by text analysis components such as Named Entity Recognition (NER) engines. ITS2.0 therefore offers a flexible palette of well-defined data categories to support the generation and consumption of content annotations by multiple processes and the translation workflow, spanning from content creation to its translation, consumption and reuse. In this sense ITS2.0 fulfills a role for the multilingual Web similar to that which the Dublin Core has played for interoperability of monolingual content publishing. In the rest of this paper we examine the content annotation techniques used in ITS2.0 separate from the semantics of the data categories it defines, with the aim of generalizing these annotations into a set of reusable patterns.

4. Generalizing ITS to Content Annotation Patterns

In considering content annotations that are suitable for deployment in existing content process chains several important principles can be derived:

a. The annotation should minimize impact on the original content so as to minimize the burden on other components in the content processing chain in handling that annotation. Impact can be assessed in terms of complexity.
b. Annotation should be well-defined in an open manner so that they can be successfully exchanged between separately implemented content processing components.
c. The mechanism for associating annotations to content should be flexible enough to accommodate different content mark-up schema, so that the processes using the annotation are not unnecessarily limited to specific content formats.
d. Consistent with point (c), annotation mechanisms should aim to be flexible enough to be associated with new content markup formats, i.e. it should be extensible
e. Consistent with points (b), (c) and (d), annotations should possess unambiguous semantics even when the mechanism for associating the annotation to content varies.
f. It should be possible to reliably remove the association of the meta-data from the content in situations where, for example, the impact of localization or personalization relate processing is not longer relevant for content reuse or other downstream processes.

The ITS approach seems to address many of the requirements, but to be able to generalize this more formally we deconstruct the various annotations into the following set of patterns. It is important to note that the specification of ITS is not based on these patterns explicitly. Therefore any attempt to build a conformant implementation should follow the ITS2.0 specification. The provisions of those specifications are written, as with any interoperability specification, to maximize the unambiguous interpretation of its provisions when building and testing a conformant implementation. In contrast, the description of patterns presented here convey some core reusable design principles underlying the ITS specifications. The aim therefore is to encourage the development of further interoperability specifications that can avail of the tried and tested interoperable content annotation solutions contained in the ITS specifications, or to extend existing ITS parsers with new data categories. Any such specification would however need to be prepared in the unambiguous manner adopted in an interoperability standard, supported by a conformance test suite.

The following annotation patterns are generalized from the established text annotations mechanism on over which consensus has been reached in the standardization of ITS 1.0 and ITS 2.0. These patterns are split between a basic set of patterns concerned with the direct annotation of textual content with attribute values, and those that offer indirect ways of associating annotation values with textual content. The pattern description describes the problem it tries to solve, the constraints under which it must be applied, the advantages of its use and where relevant explains how it is used in ITS2.0.

4.1 Direct Annotation of DOM Structured Content

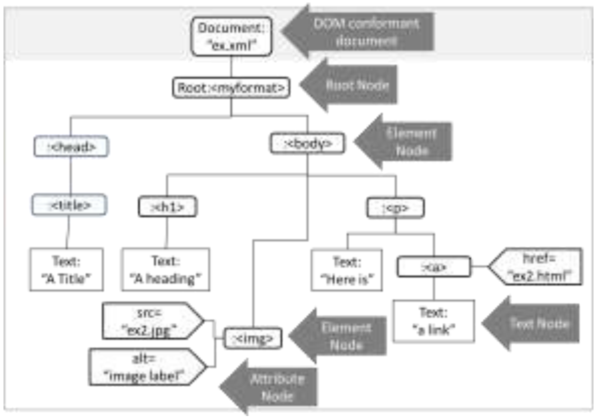


Figure 2: Example of element, attribute and text node in a DOM parse tree

P1. Annotation of Textual Content in a DOM conformant document

This specifies that 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 can be implemented by a DOM-conformant parser that enacts specific actions when detecting such a special attribute nodes associated with an element or attribute node. See figure 2 for an example of text, element and attribute node in a DOM parse tree.

This is a base pattern of the pattern language, i.e. all the other patterns rely on this one. This therefore requires that these text annotation patterns can only be applied to DOM conformant documents.

The advantage of this pattern is that by using well defined attributes to specify annotations allows these annotations to co-exist with other DOM conformant schemas in a variety of applications.

In ITS, annotating attributes are defined for XML using a specific name space and for HTML by a set of attributes with a common attribute name prefix, i.e. “its-“.

P2 Direct sub-tree annotation

In this pattern all the text nodes and text values of attribute nodes within a sub-tree of a document’s DOM representation are annotated by a well-define attribute annotating the root element of that sub tree.

The advantage of this pattern is that it allows contiguous sub-portions or a document to be easily annotated.

A constraint on this pattern is that the semantics of the annotation may not be appropriate to propagate over the text nodes and/or the text values of attribute nodes across the sub-tree. This propagating behavior therefore must be well defined for specific annotation types.

In the ITS specification, such an annotation is referred to as a local selector. The propagation of ITS annotation from a node to its sub-tree nodes is described in terms of those nodes ‘inheriting’ the annotation to the annotated sub-tree root element.

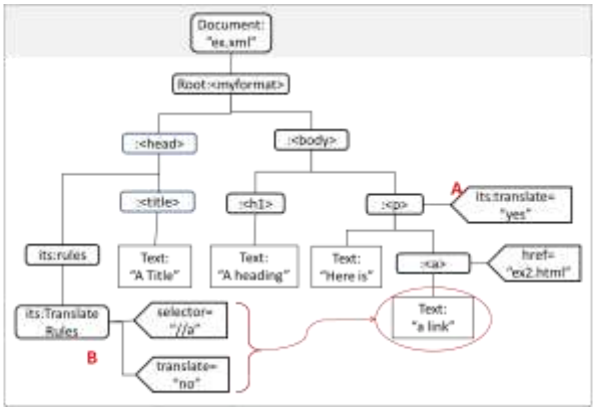


Figure 3: Examples of direct subtree annotation (A) and selector based annotation (B)

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.

A constraint of this pattern is that a new annotating element must be added to the document to house the selector-to-annotator bindings.

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. Also, as pattern P2 annotates an element it cannot be used to annotate the textual content of an attribute separately to the element which that attribute decorates. Using selector based patterns allow such attribute text values to be individually annotated.

In ITS, selector based annotations are referred to as 'global rules-based selection'. They are specified in a defined set of rule elements, which bind a specific annotation type to a specific selector. Rules elements are placed in a defined <rules> element, where multiple rules can be collected. Where rules select overlapping sets of document nodes, the order of the rule declaration is used to determine which takes precedence in parsing ITS annotations.

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.

In ITS, references to an external file with an its:rules element can be made from an Xlink hyperlink ('href') attribute from an its:rules element within the file. Rules applied in this way have a lower precedence that those declared within a document.

P5 External binding to selector-based annotation

An external definition to selector based annotation may also be bound externally to a document.

The advantage of this is that the binding can occur with no impact on the structure and content of the document.

ITS does not specify such external bind mechanisms beyond specifying that any rules applied in this manner have lower priority that those bound via an internal selector-based annotation or a references selector-based annotation. In (Ó hAirt et al 2012) we present an approach to externally binding ITS meta-data to a document in a content management systems, using the folder meta-data and multi-filing capabilities of the Content Management Information Service API standardized by OASIS (Choy et al 2010).

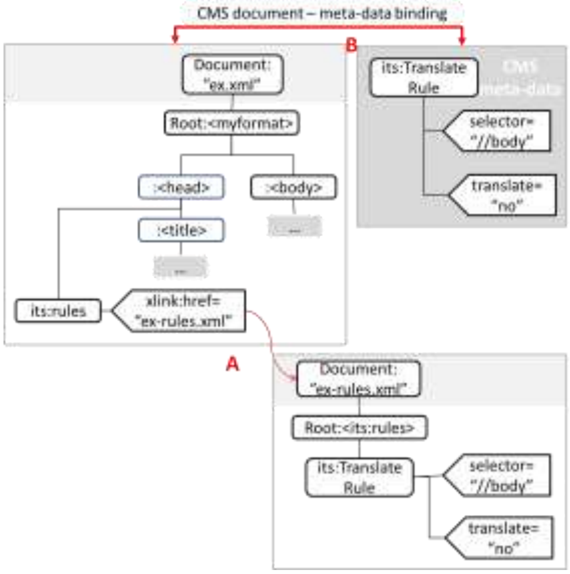


Figure 4: Examples of referenced external selector-based annotation (A) and External binding to selector-based annotation

4.2 Indirect Annotation of Structured Content

These patterns address situations where the value of annotation is not included in the attributes annotating the text, but instead the value is contained is some other meta-data that is referenced.

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 meta-data value.

The constraint is that the annotation parser must be able to access and dereference the IRI.

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. This pattern also allow for many annotated nodes to easily reference the same meta-data and it allows for that meta-data to change independently of changes to the annotated document.

Several ITS data categories contain a reference pattern data attribute, typically using the suffix 'Ref'. ITS does not specify the type of the referenced meta-data which in some cases necessitates an additional data attribute being defined to explicitly refer to a schema or classification resource.

P7 Pointer Pattern

Meta-data 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 constraint in applying this pattern is that it is appropriate to use only with the selection based annotation, i.e. it should operate as a schema level mapping, matching all selected instances of textual content to existing accompanying meta-data within a defined schema.

The advantage of this pattern is that it allows existing piece meta-data to be reused to provide interoperable textual annotation with a minimal impact on the document, thereby minimizing the necessary addition mark-up needed to achieve a new interoperable annotation

Several ITS data categories make use of this pattern, using a data attribute with a 'Pointer' suffix, the value of which must be a relative XPath selector.

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. Where multiple values need to be applied to the same text the following options can be adopted:

a) The attribute values can be specified in nested elements around the annotated text, e.g. in HTML using nested <span> elements. This has the advantage of not requiring any specialized parsing. It has the disadvantage that it adds a lot of otherwise unnecessary element mark-up to the document. This solution is not adopted explicitly in ITS.

b) The data attribute can itself have multiple values, e.g. separated by spaces. This has the advantage of being simple for single value attributes. However if the annotation requires the specification of more than one data attribute types, then a structuring convention is needed for the value, which requires its own parsing rules. These can become complex if the specification of values for all types is not mandatory. ITS adopts such a convention in the domainMapping attribute of the Domain data category. Here the multi-value is a tuple and an algorithm for parsing the values is defined. This approach also has the disadvantage that the number and size of value is limited by the maximum attribute value size.

c) Multiple annotation values may be captured as attributes of separate instances of the same element type that are collected in a special stand-off element placed elsewhere in the document and referenced by a reference pattern annotation of the text. The advantage of this pattern is that it allows straightforward DOM parsing of multiple annotations with no limit on value sizes, or the number and optionality of attribute types in a particular annotation. The disadvantage is that it introduces additional element into the annotated document. ITS2.0 implements this standoff solution for multi-annotation for the Provenance and Localization Quality Issue data categories.

P9 Annotation Meta-data

This pattern allows the annotation itself to be associated with additional meta-data. This is useful if the way in which the annotation was generated has a bearing on how it should be interpreted. It is performed by a direct sub-tree annotation whose values associate the instances of an annotation type in that sub-tree with additional meta-data.

This has the advantage of being able to annotate a large set of annotations with meta-data, without adding that meta-data to each individual annotation.

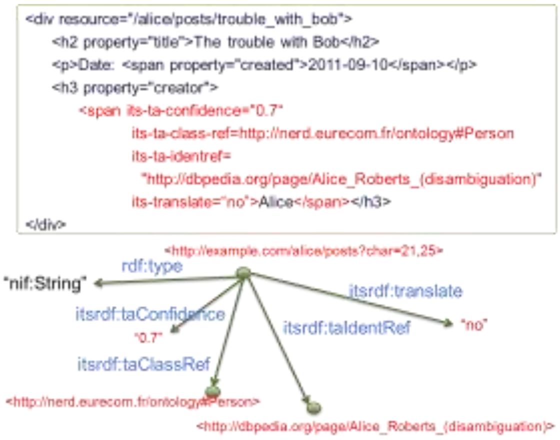
ITS 2.0 uses this pattern to associate a reference to the engine that has generated an annotation containing a confidence score with that annotation's data category. This is important since confidence scores are not comparable across engines, so identifying the engine involved is key to making use of the score. The annotation is done with the annotatorRef sub-tree annotation which can be applied to the Terminology, Text Analytics and MT Confidence data categories. This is efficient since typically all the annotation of a particular data category in a document will be performed by a single tool.

P10 External annotation of document fragments

Document may also be annotated by externally associating external meta-data with a fragment identifier in the document. The following approaches are possible:

a) An ID-based fragment IRI is used, e.g. http:://ex.xml#sect2. This is constrained however to elements with an id (or in HMTL a name) attribute defined.
b) A selector-based fragment identifier is used, using xpath e.g. http://ex.xml #xpath(/html/body[1]/h2[1]/text()[1]). This has the advantage of being able to reference any text node even if no id attribute is present. It also able to reference attribute node values. It is constrained to XML documents however, as xpath fragments are currently not defined for HTML documents.

ITS does not use either of these external fragment reference approaches directly. Instead it does specify an indirect means of externally referencing specific annotated text. This is specified as part of a mapping of ITS annotation into RDF. This involves both parsing the ITS content of document and indexing this against a version of the document where all the markup and extraneous white space has been removed and just the text characters remain. The resulting RDF model contains a string resource which uses a char format IRI, e.g. http://ex.txt#char=21-25 to identify the text segment between character count 21 and 25 inclusive, see example in figure 4. This approach can only be used with a conversion algorithm that generates such a plain text document since char fragments are not defined for XML or HTML.



```
<div resource="/alice/posts/trouble_with_bob">
    <h2 property="title">The trouble with Bob</h2>
    <p>Date: <span property="created">2011-09-10</span></p>
    <h3 property="creator">
        <span its-ta-confidence="0.7"
                its-ta-class-ref=http://nerd.eurecom.fr/ontology#Person
                its-ta-identref=
                  "http://dbpedia.org/page/Alice_Roberts_(disambiguation)"
                its-translate="no">Alice</span></h3>
</div>
```

<http://example.com/alice/posts?char=21,25>
rdf:type
"nif:String"
itsrdf:translate
itsrdf:taConfidence
"0.7"
itsrdf:taIdentRef
"no"
itsrdf:taClassRef
<http://nerd.eurecom.fr/ontology#Person>
<http://dbpedia.org/page/Alice_Roberts_(disambiguation)>

Figure 5: Example of conversion of ITS annotated content to RDF using the ITS and NIF ontologies

However, this approach does have the potential advantage of being able to specify annotations for text that is not delimited by mark-up.

5. Requirements for Personalization Annotations

The previous section shows how the wide range of annotation approaches used in ITS2.0 can be generalized into a pattern language of reusable annotation patterns. As with any pattern language patterns can be successfully applied in combination and this also is visible in the ITS2.0 specification. The benefit of this generalization is in the potential to more easily apply these annotation patterns in various combinations to the definition of new data categories.

We can therefore more easily design new set of annotation semantics and then use a process of trial implementation prototyping and consensus forming amongst concerned actors to define new set so of content processing annotation which maintain many of the benefits resulting from the design of ITS.

As a start to developing possible interoperable content annotation data categories for personalization content processing we consider the following:

- 'personalize': which indicates to downstream processes where the annotated content should or should not be personalized (analogous to 'translate' in HTML5 and ITS).
- 'slice': indicate the boundaries of a slice, perhaps with references to slicing mechanism used and a confidence score on the positioning of boundaries.
- 'domain': indicates the subject domain or domains of the content for consumption by an adaptive process, which may have an optional confidence score. In ITS, this primarily identifies existing meta-data annotation (such as HTML meta annotations) as the domain identifiers that should be used by downstream personalization processes.
- 'text analytics': this annotates content based on the output of text analysis processes to identify content for later processing. Examples of such annotation include named entity recognition or text classification. ITS has an existing annotation that can identify entity and classifying resources as URIs, accompanied by a confidence score.
- 'axes-filter': indicates the types of adaptation modes that should or should not apply to the content, e.g. language, graphical, layout, navigation, modal, phrasing, précising. ITS has a

similar data category that filters content from downstream processing based on existing BCP-47 locale codes, though here a personalization-specific coding of axes would be required.

- 'adaption-provenance': indicating what adaptation has been already applied to the content. Again there is an equivalent data category in ITS for specifying translation provenance, which can be useful in quality assurance workflows and in harvesting bi-text corpora from localization workflows using provenance parameters as a quality selection criteria. A similar role could be fulfilled for personalization, however a richer definition of agent types would be required, including: content slicer, domain annotator, text analytics annotator, indexer, filter, query rewriter, adaptive content rewriter, adaptive content composer etc. As these processes are either human driven, human checked or increasingly driven by machine learning techniques, knowing exactly which processing agents are involved in an instance of adaptation, is key in acting upon feedback received from users.
- 'adapt-script': a pointer to an executable adaptation script. This can be useful when some content is best bound directly to specific adaptation instruction that travel with the content, which may override more general processing driven by the values of other types of annotation.

These new data categories would therefore offer an abstract definition and a set of implementations, similar to ITS, enabling their implementation in HTML5, XML vocabularies and RDF data stores. However, while the evolution of ITS has been somewhat constrained by the well-established workflows already practiced in the localization industry, for personalization the pattern language presented interconnecting content and its annotating meta-data provides a well-tested starting point.

Acknowledgements

This research is supported by the European Commission as part of the MultilingualWeb-LT project (contract number 287815) and by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin. The authors would like to thank all members of the MLT-LT WG for freely giving their knowledge and guidance in support of the ideas presented in this paper.

References

Berjon, R et al (2013) HTML5, A vocabulary and associated APIs for HTML and XHTML, W3C Candidate Recommendation 6 August 2013

Buitelaar, P., Cimiano, P., Haase, P., Sintek, M., (2008) Towards Linguistically Grounded Ontologies, in The Semantic Web: Research and Applications, Lecture Notes in Computer Science Volume 5554, 2009, pp 111-125

Chiarcos, C., (2008) An ontology of linguistic annotations, LDV Forum 23(1): 1-16, 2008

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

Choy, D., Brown, A., Gur-Esh, E., McVeigh, R., Muller, F. (2010) Content Management Interoperability Services (CMIS) Version 1.0, OASIS Standard, 1 May 2010

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

Eberlein, K.J. et al (2010) Darwin Information Typing Architecture (DITA) Version 1.2, OASIS Standard, 1 December 2010

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

Hellman, S., Lehmann, J., Auer, S., Brümmer , M., (2013) Integrating NLP using Linked Data, in proc 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia

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

Koidl, K., Conlan, O., Wei, L., Sexton, A. M. (2011). Non-invasive browser based user modeling towards semantically enhanced personalization of the open web. FINA-3A: Semantic Web and Systems. Singapore

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

Levacher, K., Hynes, E., Lawless, S., O'Connor, A., & Wade, V. (2009). A Framework for content preparation to support open-corpus adaptive hypermedia. In the Proceedings of the 20th ACM Conference on Hypertext and Hypermedia. Torino, Italy.

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

Manola, F., Miller, E. (2004) RDF Primer, W3C Recommendation 10 February 2004

Ó hAirt, A., Jones, D., Finn, L., Lewis, D (2012) 'An Open Localisation Interface to CMS using OASIS Content Management Interoperability Services' at 17th LRC Internationalisation and Localisation Conference, Limerick, Ireland, accessed from http://www.localisation.ie/resources/conferences/2012/presentations/ 29th Oct 2013

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

Wade, V. (2009). Challenges for multi dimensional personalised web (Invited Keynote). User Modeling, Adaptation, and Personalization Conference (UMAP 2009). Trento, Italy

Windhouwer & Wright (2012) Linking to linguistic data categories in ISOcat, , in proc Linked Data in Linguistics 2012