A Semantic Model for Integrated Content Management, Localisation and Language Technology Processing

Dominic Jones[1], Alexander O'Connor[1], Yalemisew M. Abgaz[2], David Lewis[1]

[1 & 2] Centre for Next Generation Localisation
[1]Knowledge and Data Engineering Group,
[1]School of Computer Science and Statistics, Trinity College Dublin, Ireland
{Dominic.Jones, Alex.OConnor, Dave.Lewis}@scss.tcd.ie

[2]School of Computing, Dublin City University, Dublin, Ireland
[2] Yabgaz@computing.dcu.ie

Abstract. Providers of products and services are faced with the dual challenge of supporting the languages and individual needs of the global customer while also accommodating the increasing relevance of user-generated content. As a result, the content and localisation industries must now evolve rapidly from manually processing predicable content which arrives in large jobs to the highly automated processing of streams of fast moving, heterogeneous and unpredictable content. This requires a new generation of digital content management technologies that combine the agile flow of content from developers to localisers and consumers with the data-driven language technologies needed to handle the volume of content required to feed the demands of global markets. Data-driven technologies such as statistical machine translation, cross-lingual information retrieval, sentiment analysis and automatic speech recognition, all rely on high quality training content, which in turn must be continually harvested based on the human quality judgments made across the end-to-end content processing flow. This paper presents the motivation, approach and initial semantic models of a collection of research demonstrators where they represent a part of, or a step towards, documenting in a semantic model the multi-lingual semantic web.

Keywords: Multilingual Web, Content Management, Localisation, Language Technology and Interoperability

1. Introduction

To engage successfully with global markets, enterprises increasingly need to manage fast moving streams of multi-lingual content from both within the enterprise and from the wider user communities with which they wish to engage. For example, modern software markets are increasingly dominated by larger numbers of fine-grained applications, e.g. smartphone and social network 'apps' or Software-as-a-Service (SaaS) offerings, that feature high frequency/ "perpetual beta" release cycles. For these products, technical documentation increasingly adopts the form of FAQs, blogs and wikis that grow with significant input from users and customer support staff as features and problems are discovered and solutions documented. Even when technical manuals are provided (and localised into different target markets), users increasingly find more direct solutions to problems on user-generated question & answer sites, which may then themselves, merit localisation based on demand.

Managing this multilingual content stream requires seamless integration of content management systems, the localisation chain and data-driven natural language technologies. Cost, timeliness and quality trade-offs need to be actively managed for different content flows with a far greater level of flexibility and automation than that that has been achieved previously to cover the range of multilingual content generation and consumption paths, e.g. from online documentation, to Q-A fora and micro-blog and RSS feeds.

The Centre for Next Generation Localisation (CNGL) specifically consists of over 100 researchers from academia and industry who conduct research into the integration of natural language technologies, localisation and digital content management required in addressing these challenges. Researchers work collaboratively in CNGL to produce a range of technical demonstrators that integrate multiple forms of multilingual content. A key challenge associated with such large-scale research systems integration is the need for researchers to collaborate and for software to interoperate. However, the components are derived from a number of different research and industrial communities, where either meta-data was not formally defined or was specified from a fragmented set of standards or industry specifications. CNGL therefore established a meta-data group (MDG) to concentrate and integrate the meta-data expertise from these different research areas, including statistical machine translation and text analytics research, adaptive content and personalisation research and localisation workflow and interoperability. To address the universal trend of content to be web based and to offer a well-supported, community-neutral approach to semantic modelling, the standardised languages of the W3C Semantic Web initiative were used. This allowed multiple existing meta-data standards and component meta-data requirements to be incorporated into a single model thereby demonstrating the interrelation and utility of such interlinked meta-data. This paper presents the initial process for the development of such semantic models for interoperation within a large software research project, such as the CNGL. Future research will see these models are deployed and evaluated in multiple industrial settings to establish their broader applicability in the field of localisation and digital content management. This paper only provides a discussion around forming these initial models.

2. Approach

There is an increasing focus on the interoperability of Content Management Systems (CMS), the tools in the localisation chains and the emerging array of language technologies offered as services. In the localisation industry there are many different platforms, favoured by various Language Service Providers (LSPs) for translation projects. Some are in-house deployments, some open-source (such as GlobalSight) and some purchasable from third parties (such as SDL WorldServer). However each platform comes with its own nuances, meta-data specification and translation workflow management tools. For example within Digital Content Management multiple approaches to storing meta-data apply ranging from storing simple XML content as mark-up (from both authors and consumers) through to complete Ontological models. In Machine Translation meta-data appears in the many forms from terminological mark-up, workflow management meta-data and Translation Memory (TM) that allows previous translations to be recycled in the Machine Translation (MT) process. As has been shown complex meta-data standards occur in many places across both the multi-lingual semantic web and localisation industry.

CNGL develops a large number of distinct "demo" systems integrating different aspects of content management, localization and language technology integration.

Figure 1 summarises how the current range of integrated systems in these categories map onto the NGL Process Map. Semantic models, consisting of content and service models are developed by synthesizing a seed model from an analysis conducted across the range of these integrated systems. This model has then been mapped back onto revisions of the integrated system to assess the degree to which the model can accommodate the natural evolution of these systems as they evolve in response to specific industry requirements and technological advances.

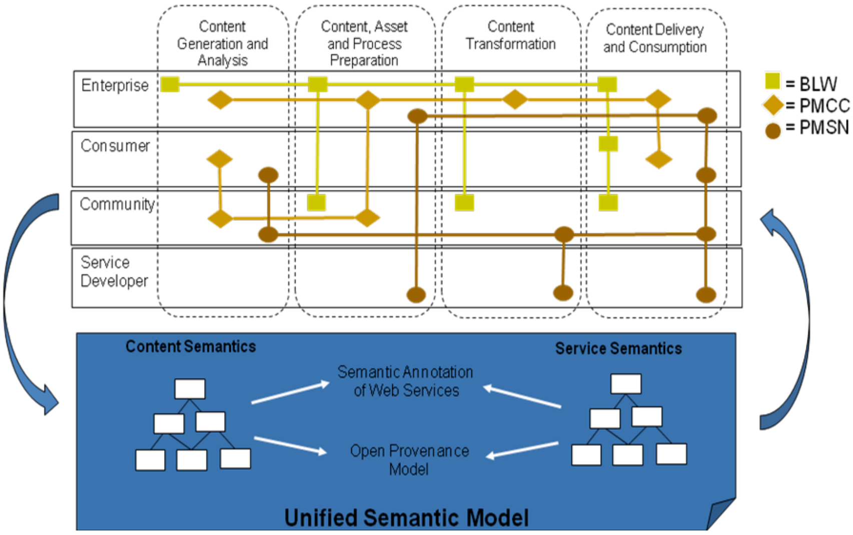

Figure 1: NGL Process Map used for Organizing Semantic Modeling

The availability of such a large, motivated cohort of collaborating researchers presented a range of distinct views and expertise and allowed us to gather input for and define common models of both the content and services offered across Localisation, Content-management and Integrated Language Technologies. The demonstration systems used address use scenarios in the areas of: Bulk Localisation Workflow (BLW); Personalised Multilingual Customer Care (PMCC) and Personalised Multilingual Social Networking (PMSN). To develop a shared understanding of this modelling, demonstration systems workflows and the components implementing specific activities are overlaid onto a Next Generation

Localisation (NGL) Process Map (Figure 1) that identified different stakeholders and abstractions of the major process areas involved in content generation, its preparation for manual and machine processing, the translation, such as translation and personalisation, resulting from this processing and the delivery and consumption (including rating and feedback) by consumers. Semantic models, the content being processed and the services offered by the components doing these processing were then developed by synthesizing a seed model from an analysis conducted across the range of integrated systems and making continual iterations to this model.

However, the breadth of input also put into sharp focus the degree of change these models would be subject to, resulting from continual advances in the research components used, we therefore applied Ontological Evolution techniques to create a common evolving model that is representative of this set of changing viewpoints and that is extensible for use by others in the future. This semantic model has therefore been mapped back onto revisions of the integrated demonstration systems to assess the degree to which the model can accommodate the natural evolution of these systems (which have undergone four coordinated iterations over the last three years) as they change in response to specific industry requirements and technological advances.

3. Challenges for Multilingual Web Meta-data Interoperability

The purpose of building common semantic models is driven by the fact that standardisation efforts in support of content management and localisation processes and tool interoperability are somewhat fragmented between the Organization for the Advancement of Structured Information Standards (OASIS), the Open Standards for Container/Content Allowing Re-use (OSCAR) operated by the (now defunct) Localisation Industry Standards Association (LISA) and the W3C. Most current standards are in the form of XML document schema designed to support hand-over of specific content and meta-data between different tools through import and export functions using these different file formats. This occurs primarily when a translation job is handed off from a content developer to an Language Service Provider (LSP) via their respective Translation Management System (TMS) tools and onwards to an individual translator's CAT tool, possibly via a translation agency [1].

For example LISA's Translation Memory Exchange (TMX) [15] standard is one of the most widely used and supports the hand-off of the key linguistic assets currently used in the translation process. This is typically from Translation Memory (TM) repositories to Computer Aided Translation (CAT) tools, where content is leveraged, and back again as quality-check translations that are added to the TM. Similarly, LISA defined Term Base Exchange (TBX) [16] which is an XML format for exchanging terminological information, aligned with description categories defined by ISO/TC37 on Terminology and other language and content resources.

The XML Localization Interchange File Format (XLIFF) [14] from OASIS offers a standard way of passing a live translation job (combined in an envelope of content and meta-data between different tools) handling: the extraction and segmentation of translatable text; its translation by human, TM leverage or Machine Translation (MT); translation review and re-integration of target language content with formatting skeleton files for publication. The XLIFF schema suffers however from a high level of optional elements as well as common use of user-generated extensions resulting in difficulties in its use for blind interchange of localisation meta-data between tools [2]. In addition to this XLIFF does not support the passage of process meta-data from either content authoring processes (including the application of controlled language) to the localisation process, nor from the localisation process to the TM and terminology maintenance processes, so that meta-data potentially useful for future TM leverage or TM and terminology use in Statistical Machine Translation (SMT) training is lost.

Another OASIS TC, on an Open Architecture for XML Authoring and Localization (OAXAL) [18], has articulated a reference model that proposes using the XML Text Memory schema defined by LISA (xml:tm) [19] to provide better end-to-end management of localised content. By managing the segmentation of source content within the content management system, it provides both author memory to assist changes in the source termed 'translation memory', and maintains the link between source and target content such that changes to either can be managed and also used to more effectively build new TMs. However, only a high-level integration framework has been outlined to date and considerable additional work toward alignment with other standards would be required for a workable solution. In summary the challenges faced in supporting integration of content processing across the content management and localisation industries are seen as:

- The tendency towards tool vendor lock-in, in terms of both Translation Management Systems and exchange formation, and the associated lack of a strong independent tool vendor market supported by widely adopted common standards.
- The need to migrate localisation processes to integrate with increasingly web based, multi-media and multi-modal content that can be personalised to individual user and community needs.
- The need to apply localisation and personalisation processes to user generated content and live social media while also actively leveraging the wisdom of the crowd in accelerating those processes. More and more content is originating from rapidly changing sources of content, involving the user in the translation and post-editing of such content is beneficial to the user (being provided access to content in their native tongue) and to the producer in terms of reduced costs associated with translation. Where reduced costs, increased speed, free marketing and delivering of quality assurance from the users themselves form a new approach for companies to translate content.
- The lack of interoperability standards that span the Next Generation Localisation (NGL) problem space and the lack of maturity of many of the standards that are presented in various sub-domains.

In order to begin to address these challenges integration in CNGL is based around the adoption of Service Oriented Architectures (SOA). However, the wide and diverse scope of CNGL requires strong common models so that meta-data can be easily shared and services integrated without expensive mismatches in assumptions about data and state and with clear processing expectations for the use of services.

Although localisation tool vendors are already attempting to support interoperability through exposing web service APIs to the functionality they offer, these interfaces are still largely proprietary, though there have been attempts to use XLIFF and TMX file format as input/output message payloads. Service-oriented integration of language technology has received some attention also by the research community, including: the Japanese-funded LanguageGrid project that provided a platform for the integration of arbitrary compositions in language resources [3], the CNGL which is applying service oriented techniques to integrating language and digital content management technologies with localisation workflows [4] and more recently the Panacea FP7 [5] project which is investigating service compositions for language resource processing pipelines.

However the document-centric nature of existing localisation standards means that their use in web service interfaces for localisation has been ad hoc and lacked any shared conceptual or common conformance framework needed to support seamless integration. In contrast, our semantic model forms a shared conceptual view of services and content that has been derived and validated through a set of software integrations. By providing a core common data type model with a well defined conceptual basis for service, developers can define their interfaces more precisely as contracts, with an easily reached shared understanding of precondition states and processing expectation involved in invoking a given service. At the same time, however, we do not aim to supplant existing interoperability standards where they exist, nor do we intend to be a source of new interoperability standards. Instead we aim to provide a minimal common model for data types that can be exchanged via NGL services, while leaving it to individual service developers to use this type set to define their service interfaces in their particular exchange format.

4. Semantic Model

As the common focus of integration was the processing of digital content, a content-based semantic model was developed that incorporated a content processing service model **Figure 2** and a processed (i.e. managed) content data model Figure 3. Both of these models have been developed using the Resource Description Framework (RDF) [6] and ontological principles from the Semantic Web to provide:

- Flexibility in defining types.
- Extensibility through class specialisation.
- Operational persistence with explicit meta-data in the form of triple stores.
- Future support for open linked data approaches to content processing.

This approach has provided a more flexible and extensible mechanism for defining data types that can be exchanged between what we expect to be a large and dynamic set of NGL services. In addition the semantic models also help to simplify the design and refinement of content processing workflows in the CNGL research space by interlinking concepts from existing models and existing standards including Content management including DocBook [12], DITA[13], HTML, XML and Localisation standards including: XLIFF [14], TMX [15], TBX [16], LCX [17], OAXAL [18] and xml:tm [19]. The problem with such standards is that they often are deployed or utilised in separate deployments or installations crossing over one another without a common theme to their use where different platforms utilise different standards making their interoperation hard to achieve. In the CNGL the aim is to make the semantic model open, since to reduce industry interoperability costs, a common model must be widely adopted by other system integrators and tool vendors and this model must grow and evolve to adapt as new opportunities are exposed by third party usage and their resulting feedback. We aim to improve the quality of the semantic model by treating CNGL's broad range of demos as a unique Interoperability Laboratory providing, revising and reviewing semantic models sourced from a seed model and synthesised from an analysis of early iterations of integrated systems. Future research will see this model deployed in an industrial setting where the users of the localisation standards presented can evaluate how well such a common model applies in their particular workflow.

The seed Semantic Model is broken down into two core parts, these are: **Service:** High level classifications of the different services covered based on their content processing features **and Content:** The core processed content model that we use to record the content transformations of various types (including use of content as MT training data) delivered by either activities from the business model or services. These models provide an open mechanism for defining common meta-data from existing software development, modelling tools and persistent data-stores. It therefore offers a practical mechanism for defining a minimal core set of data types that can then be composed or extended as new services are defined and new application requirements arise. A benefit of such an approach towards semantic modelling is that it supports the development of content processing management logs, which due to the emerging nature of the services and applications involved, must be more exploratory in nature, while operational, configuration and tuning requirements are established for data-driven NLP technologies.

During the process of model evolution a parent-child relationship is formed between classes of content, so for example a sub-class of "GenerateContent" is "AuthorText" following a more specific or less specific formation of relationships. Properties are assigned to classes as data-type relationships for example "AuthorText" may have a property "WordCount = int" where users are free in their adaptation of the models to add as many sub-classes, super-classes or class structure changes as they wish. In the Content Model users are able to add as many properties as they wish. All of the changes to all of the models are integrated in one single collaborative session where the meta-data group debates, integrates and records the changes made to the initial seed models. The versions presented in this paper are the initial seed models of both content and services with version 1 soon to be released.

4.1. Service Model

The Service Semantic model, shown in **Figure 2**, is based around the assumption that all services or process, create or consume content in some form or another. As RDF allows multiple inheritances defining new class types, the schema is defined to allow integration of fundamental aspects of content processing services to define a wide range of services. The core upper level service types include:

- **GenerateContent**: the creation of content by human users.
- **TransformContent**: the transformation of content from one human understandable form to another, including translation, text to speech and content personalised or adaption for delivery to a particular user.
- **AnnotateContent**: where additional meta-data is associated with content.
- **ProcessGroupContent**: where operations on **sets** of content are represented.
- **CreateService**: allowing the creation of a service to result from a processing chain of other services, therefore allowing the configuration and training of an SMT engine or other data-driven components, or adaptive composition of services to be captured.

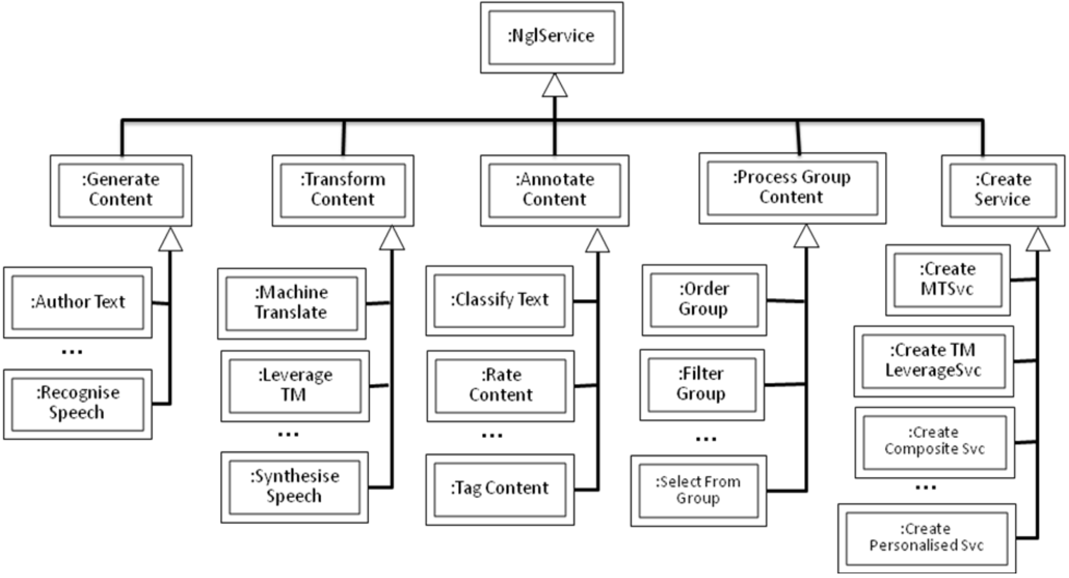

Figure 2: Current "Seed" Upper Layer Taxonomy of Service Types

4.2. Content Model

Similar to the Service Model, the Content Model, shown in Figure 3, aims to provide an extensible fine-grained schema of types from which the input and output of services that can be modelled. The core type in the content model is ManagedContent, which indicates that we are only interested in content that is subject to some form of management or monitoring. The following key content processing content types are then identified:

- **GeneratedContent**: content produced by a person.
- **AnalysedContent**: content that has been analysed prior to making some decisions on further processing, such as user adaption.
- **PreparedContent**: content that has been altered for further processing.
- **LocalisedContent**: content that has been subject to the localisation processes.

Other seed subclasses have been identified to differentiate content that has been personalised, published or presented to a user or been discovered via information retrieval. Further orthogonal subclasses differentiate: the manual and automatic processing of content, content that is managed as an asset, e.g. a linguistic resource, as well as utility types indicating the content is grouped, time-stamped and serialised as a file, has had its elements counted or some intellectual property right asserted over it. In this seed structure it is important to note that “PreparedForLocalisation” and “PreparedForAdapation” are different in the following way: Localised content is content which is translated and personalised for delivery to a user, Adapted content is only content which is personalised and may not specifically have been translated before delivery to a user. Also important to note is that in terms of both the content and service models these change as the crowd of users adapt them to include the services and content types offered through their research. For example the content model does not include any user generated or live social media content types, however once deployed and adapted by users using a crowd-sourced approach such additions are deemed more likely.

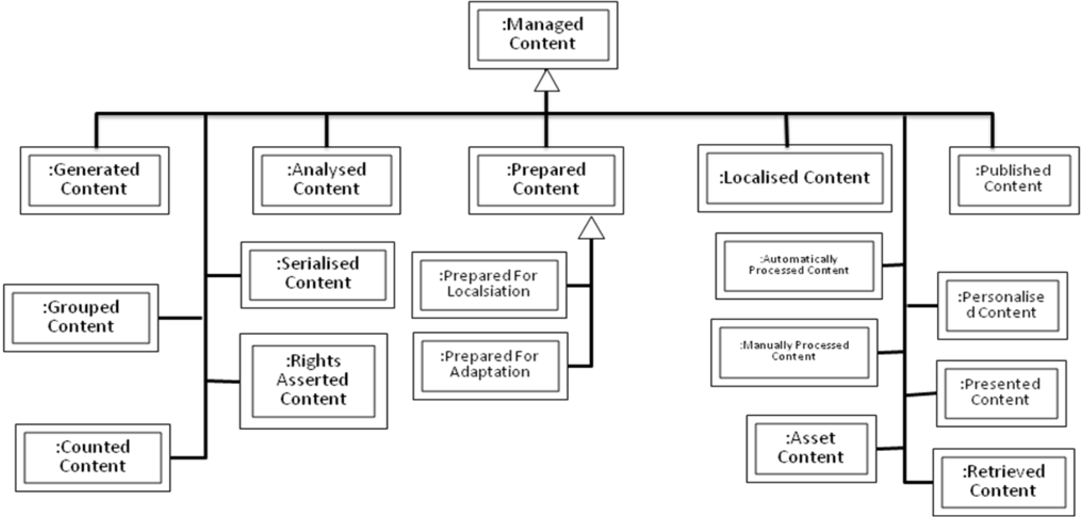

Figure 3: Current "Seed" Upper Layer Taxonomy of Content Types

5. Ontology and Semantic Evolution

The models presented here have been captured as an ontology using RDF. An Ontology or semantic model is a specification of a shared conceptualization of a domain [7] and serve to capture domain knowledge in a generic way providing a commonly agreed understanding and conceptualization of knowledge elements within that domain. They provide common ground for understanding, conceptualization, representation and interpretation of domain concepts uniformly across different systems, languages and formats. However, as discussed above, the dynamics of the given ontology often demand changes to application requirements that may be fulfilled only by changing/evolving these principal ontologies [8].

In [9 and 10] ontology evolution is defined as: “The timely adaptation of an ontology to changed patterns of usage in the ontology-based application, as well as the consistent management and propagation of these changes to dependent elements.” Ontology evolution takes place when an ontology management system facilitates the modification of an ontology by applying the proposed changes to the model and by ensuring its subsequent consistency. The need for ontology evolution is directly related to the changes that occur in the domain area or in the business environment and reasons for change include conceptualization, representation or specification of the domain. State-of-the-art ontology evolution process has six defined phases [8,9,10]. The first phase, “change capture”, focuses on investigation and capturing changes to the domain such as new concepts, out-dated concepts or updated concepts. The second phase is “change representation”. Captured changes are represented using atomic or composite change operations. The “semantics of change” phase deals with the effects of the changes that have been requested, checking the effects of the changes if implemented in the ontology [11]. The main task of the next phase, “change propagation”, is to confirm that dependent ontologies and tools are still functioning properly and that no inconsistencies are introduced by changes to the ontology. The “change implementation” phase focuses on implementing changes to the ontology and keeping a record of these changes for undo or redo purposes. Finally the change validation phase validates the change and makes it publicly available after resolving inconsistencies if there are any.

In a collaborative environment such as CNGL, achieving a common semantic model requires a continuous evolution of models. Users expect the model to support their changing requirements and thus leave the model in a continuous evolution. Researchers applying their technical demonstrators to the current model of services and content adapt the models presented in this paper and these changes are incorporated into the seed models producing iterative versions of common, CNGL wide, agreed semantic models, the first full versions soon to be published.

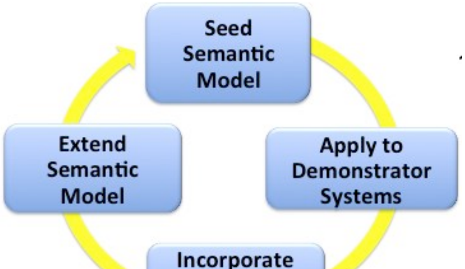
Seed Semantic Model
Apply to Demonstrator Systems
Incorporate Feedback from Demo teams
Extend Semantic Model

Figure 4: Approach taken to developing semantic models

The approach taken, shown in Figure 4, to populating the semantic models, from the collection of demonstrator systems, across the CNGL, involved producing a seed semantic model applying the model to each individual demo team. Incorporating feedback from the collection of demo teams into the seed models and extending the semantic model through a process of continual evolution using a panel of domain experts (the meta-data group) to enforce chances. In this approach, the need for evolving the semantic model comes directly from the demonstrators.

In the process of building the model, new services and content types emerged from the demo teams. To incorporate these new services and content types, it is essential for the models to evolve. This evolution occurs whilst incorporating feedback from seventeen demo teams through an interview approach to data collection. A total of 43 change requests were made to the NGL service model (12 changes) and NGL content model (31 changes). These requests included the addition of new content types which were not captured in the seed model, specialization of existing concepts into two or more types, generalization of categories to a single type and renaming of existing types. There were no requests made for deletion of a given content or service type, however in the process of implementing the above changes, for example renaming, there were deletions introduced. In addition to working on the semantic models of services and content properties were incorporated into the content model 44 of which were identified. Some example of additions made in populating the models include:

Requested content changes:

- Add subclass (user feedback, managed content)
- Add subclass(user opinion, user feedback)
- Add subclass(user input, user feedback)

Requested service changes:

- Add subclass(rank TM, process group content)
- Add subclass(compare content, process Group content)
- Add subclass(classify audio, annotate content)

Requested properties:

- Add Property (source Language)
- Add Property (input Language)
- Add property (number Of Sentences)

6. Conclusions

Localization is moving more and more towards Content Management in terms of the process pipeline apparent in taking a piece of content and turning it from source language to destination. As the breadth and depth of the content being localized increases there becomes a more apparent need to move towards a common semantic model of the content being processed. This model needs to be both human understandable but also normative, and for this reason it is argued that RDF is a suitable candidate for building models of a rapidly expanding content set. This paper has presented an approach currently being utilized for building common semantic models of the services and content types from a number of demonstrator systems focusing on localization, translation, natural language processing technologies and digital content management. It is argued that such a collaborative approach to gathering and defining semantic models of services and content provision is vital in the future of the multi-lingual semantic web as it attempts to span the different related research and industrial communities.

7. Acknowledgements

In acknowledgement this research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin and Dublin City University.

8. References

1. Gough, J. “A troubled relationship: the compatibility of CAT tools”, TAUS technology article, downloaded from http://www.translationautomation.com/technology/a-troubled-relationship-the-compatibility-of-cat-tools.html 28 Dec (2010)
2. Mateos, J “ A Web Services Approach to Software Localisation. Bringing Software Localisation Tools from the Desktop to the Cloud”, Trinity College Dublin, Computer Science Technical Report TCD-CS-2010-25 (MSc Dissertation), 20 October 2010, https://www.scss.tcd.ie/publications/tech-reports/reports.10/TCD-CS-2010-25.pdf -01/11
3. Inaba, R., Murakami, Y., Nadamoto, A., Ishida T. “Multilingual Communication Support Using the Language Grid” Intercultural Collaboration, LNCS 4568, Springer, Aug 2007
4. David Lewis, Stephen Curran, Dominic Jones, John Moran, Kevin Feeney (2010). An Open Service Framework for Next Generation Localisation. Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation, Malta, May 2010.
5. A. Toral, "Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation." (2011)
6. O. Lassila and R. R. Swick, "Resource description framework (RDF) model and syntax," *World Wide Web Consortium, http://www.w3.org/TR/WD-rdf-syntax*
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge. Acquisition. 5(2) (1993) 199–220
8. A., Stojanovic, N., Studer, R.: Ontology evolution as reconfiguration design problem solving. Proceedings of the 2nd international conference on Knowledge capture (2003)
9. Zablith, F.: Dynamic ontology evolution. International Semantic Web Conference (ISWC) Doctoral Consortium, Karlsruhe, Germany (2008)
10. Stojanovic, L., Maedche, A., Motik, B., Stojanovic, N.: User-driven ontology evolution management. Lecture Notes in Computer Science. 6(4) (2002) 285–300
11. Abgaz, Y., Javed, M., Pahl, C.: Empirical analysis of impacts of instance-driven changes in ontologies. In: On the Move to Meaningful Internet Systems: OTM 2010.
12. DockBook, http://www.docbook.org/
13. Darwin Information Typing Architecture (DITA) http://dita.xml.org/ Accessed - 09/11
14. XML Localisation Interchange File Format (XLIFF) http://www.oasis-open.org/committees/xliff/ Accessed - 09/11
15. Translation Memory eXchange (TMX) http://www.lisa.org/fileadminstandards/tmx1.4/tmx.htm Accessed - 09/11
16. TermBase eXchange (TBX) http://www.ttt.org/tbx/ Accessed - 09/11
17. Localisation Content Exchange File Format (LCX) http://www.localisation.ie/xliff/resources/presentations/lcx-xliff.pdf Accessed - 09/11
18. Open Architecture for XML Authoring and Localization (OAXAL) http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=oaxal Accessed - 09/11
19. XML-based Text Memory (Xml:tm) http://www.infomanagementcenter.com/enewsletter/200608/fifth.htm Accessed - 09/11