On Using Linked Data for Language Resource Sharing in the Long Tail of the Localisation Market

David Lewis, Alexander O'Connor
Centre for Next Generation Localisation
Knowledge and Data Engineering Group, Trinity College Dublin
E-mail: dave.lewis@scss.tcd.ie alexander.oconnor@scss.tcd.ie

Andrzej Zydroń
XTM International
E-mail: azydron@xtm-intl.com

Gerd Sjögren
Interverbum Technology
E-mail: Gerd.Sjogren@interverbumtech.com

Rahzeb Choudhury
TAUS
E-mail: rahzeb@translationautomation.com

Abstract

Innovations in localisation have focused on the collection and leverage of language resources. However, smaller localisation clients and Language Service Providers are poorly positioned to exploit the benefits of language resource reuse in comparison to larger companies. Their low throughput of localised content means they have little opportunity to amass significant resources, such as Translation memories and Terminology databases, to reuse between jobs or to train statistical machine translation engines tailored to their domain specialisms and language pairs. We propose addressing this disadvantage via the sharing and pooling of language resources. However, the current localisation standards do not support multiparty sharing, are not well integrated with emerging language resource standards and do not address key requirements in determining ownership and license terms for resources. We survey standards and research in the area of Localisation, Language Resources and Language Technologies to leverage existing localisation standards via Linked Data methodologies. This points to the potential of using semantic representation of existing data models for localisation workflow metadata, terminology, parallel text, provenance and access control, which we illustrate with an RDF example.

Keywords: Localisation, Linked Data, Language resource sharing, RDF

1. Introduction

The localisation industry consists of content generating enterprises that, acting as localisation clients, engage Language Service Providers (LSPs) to translate source content. In recent decades, the main technological innovations to bring productivity improvements to the localisation industry have involved the collection and reuse of language resources. Specifically these are Term-bases, which are multilingual glossaries that improve consistency in both authoring and translation of terms, and translation memories, which are databases of previously translated sentences that assist translators in translating identical or similar sentences, phrases or terms. More recently, Translation Memories and Term-Bases are being reused by LSPs as a source of Parallel Text that provide good quality training corpora for Statistical Machine Translation (SMT) engines.

Despite these trends, smaller localisation clients and LSPs are poorly positioned to exploit the benefits of language resource reuse in comparison to larger companies. Their low throughput of localised content means they have little opportunity to amass significant term-bases and translation memories as assets to reuse between jobs or to train SMT engines tailored to their domain specialisms and language pairs. This is compounded by the lack of overhead capacity which is needed to maintain these resources and their reuse potential.

The potential for the localisation industry to benefit from pooling and sharing language resources has already been recognised. For example, the TAUS Data Association (TDA www.tausdata.org), has pioneered language resource sharing that offers localisation clients and LSPs the opportunity to access pooled translation memories, primarily to train SMT engines, in volumes unavailable to most of them previously. However, the need to bootstrap this process with large volumes of translation memory data has initially favoured engagement from larger enterprises and prompted a centralised sharing approach

We propose a potentially more decentralised approach to sharing language resources for the localisation industry on the web as linked data, i.e. fine grained, inter-linked data elements accessible via individual URLs. This approach allows resource consumers to search and filter over distributed sources at different levels of granularity

by using meta-data languages and queries. This builds on the W3C's Semantic Web Standards; Resource Description Framework (RDF) and the Simple Protocol and RDF Query Language (SPARQL). These standards allow web content and web data (i.e. deep web content) to be interlinked in a decentralised and distributed manner, while remaining discoverable by sharing partners at any time through SPARQL queries. This approach may offer agility to third parties in adding value to existing shared resources via links from elements capturing that value to the original resource. This allows both the value-add element and the original resource to be easily tracked across localisation workflows and their reuse audited. This approach has the potential to be much more open and extensible than existing language resource sharing schemes. It would however require flexible access control and reuse auditing features to support innovation in novel sharing schemes that may be tailored to specific market niches, e.g. medical terminology. The result would be an interlinked web of language data resources and associated provenance meta-data which has been generated during resource creation and reuse in the localisation process. We refer to this as *Linked Language and Localisation Data*, or more concisely *L3Data*.

As previously reported (Lewis et al 2012), we have developed a simple localisation crowd-sourcing application based on L3Data handling principles as a basis for exploring future L3Data sharing scenarios. This paper surveys existing approaches to linguistic data sharing from the linked data community, the localisation industry and the language resource sharing community associated with natural language processing research. In each case we examine how these approaches could be extended to support the proposed vision of L3Data sharing using linked data, and provide an example based on our current implementation.

2. Resource Sharing via Linked Data

In general terms, the Linked Data architecture of the web consists of sets of facts stored as RDF triples. These amount to assertions about particular individual entities (such as a particular document) and its properties (such as an author). RDF can adopt a variety of schemata to represent and classify the properties and individuals, for example the Document Class, or the Author. The schemata provide a method for accessing data across different repositories. The triples can be queried using SPARQL, a pattern matching query language, which permits complex queries to be resolved against the facts. This creates a relatively light-weight, highly-interoperable, standards-based model for sharing data across the web.

The two key innovations in Linked Data are the ability to interlink data between different knowledge repositories, and the ability to represent facts in lightweight structured form based on semantic web standards. Beyond these language level standards, Linked Data repositories make maximum reuse of existing published vocabularies (such as document 'Subjects' from Dublin Core - http://dublincore.org/, or 'Person' from Friend of a Friend - http://www.foaf-project.org/). Consensus within domain communities on appropriate RDF vocabularies is primarily driven by uptake of vocabularies across the volume of published data rather than prescription by standards bodies.

The principal sharing application of Linked Data is the Linked Open Data (LOD) cloud, seeded from DBPedia (Auer et al 2007), which leveraged structured knowledge from Wikipedia. The LOD cloud now contains over 31 billlion triples and over 500 Million links between sources provided from 295 datasets, the majority being governmental (Bizer et al 2011). Though the vast majority of current datasets in the LOD cloud are in English, there is strong standards support for multilingual linked data through its use of Unicode, RDF's element-level language tagging and International Web Resource identifiers.

Relevant best practice in linked data sharing suggests use of the Open Provenance Model (Moreau et al 2008) for recording changes to shared data; of licensing annotations to record assertion of ownership, copyright and use permissions; and of the use of the VoID vocabulary (Alexander et al 2009) to aid discovery of a dataset, its API and SPARQL end points.

One challenge in supporting decentralised language data resource sharing in a commercial setting is the control of access to those resources. This is required to control exposure of confidential or copyrighted content and to prevent the fear of free riding by competitors from deterring sharing contributions from SMEs. Solutions for access control of RDF triple stores have involved use of policy based management techniques (Abel et al 2007), use of semantic web rule languages (Muhleisen et al 2010), and access control vocabularies for data sets (Villata et al 2011).

3. Language Data Resource Sharing in the Localisation Industry

A majority of organisations have a need for terminology management to support content creation and translation management (Wright & Budin 1997). A major obstacle to smaller organisations engaging in terminology initiatives is the cost of building and maintaining term databases, including the cost of accessing the necessary tools. The standardised expression of terminological and other lexical resources has been addressed by ISO Technical Committee resulting in an extensible Language Mark-up Framework (Francopoulo et al 2006) of data categories that can capture different lexical properties of a term. While this has proved too complex to adopts in its entirety in the localisation sector, ISO/TC37 produced a more tractable subset encoded in XML, TermBase eXchange (TBX), which is now supported in some localisation industry tools. There are increasing numbers of online

dictionaries that offer searches on terms their translations into different languages, e.g. glosbe (http://glosbe.com), however these are often closed services, with few offering opportunities to contribute to or download the dictionary.

Translation Management tools are key to delivering increased localisation productivity. Translation is a highly collaborative process involving project managers, translators, reviewers and correctors (Karamanis et al 2010). Translation management also involves sharing resources such as terminology, translation memory and the file being translated. However, this language resource sharing largely only occurs downstream along localisation workflows. The downstream sharing of clients Translation Memories plays an important part in job pricing, as the number of 'in context exact' matches as well as fuzzy matches between the job content and the translation memory impacts the pricing negotiated with the LSPs. However, only completed translation are returned upstream, with few incentives in place to return error reports or quality improvement to received translation memories, or the term bases that often accompany them (Lewis et al 2009). This downstream-centric, push-based approach to language resource exchange is reflected in the standards that have been developed for the localisation industry, namely TermBase Exchange (TBX) for terminology, Translation Memory Exchange (TMX 2005) and XML Localisation Interchange File Format (XLIFF 2007) . These are typically implemented in job-level tool import/export functions that preclude any fine-grained round-trip consistency management as the language resources or the content are updated over time. The OASIS OASIS Open Architecture for XML Authoring and Localization Reference Model (OAXAL - http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=oaxal) technical committee, proposes mechanisms for linking from source web content to term-base entries (using the Termlink XML schema) and segment-level translations (using the XML Text Memory, 'xml:tm' schema). Currently this linking points to accompanying term-bases and translation memories, however, the use of explicit links could be readily extended to URLs to linked data stores. The Multilingual Information Framework (MLIF) (Cruz-Lara 2004), being developed under ISO TC37, offers a model that forms links between TBX, TMX, XLIFF and the W3C's International Tag Set (Lieske & Sasaki 2007), through a set of interlinked data categories. The RDF mapping already explored under the LMF initiative, therefore provide an as yet unexplored route to readily implemented MLIF as a linked data vocabulary

Using linked data to enable a pull-based approach to exchange language resources addresses the problem in the current push-model with language resource staleness, such that active translation processes can be immediately notified to updates to term-bases and translation memories. By adopting the RDF vocabulary for the Open Provenance Model (OPM) as a standard way to record changes to linked data, updates to the term bases and translation memories by translators may be logged, audited and potentially remunerated, thereby recognising the contribution and encouraging further updates. Further, by using OPM to record the localisation steps performed on individual segments, a fine grained log of quality assurance data is amassed, adding value to term-bases and translation memories as reusable resources.

4. Language Resource Sharing for Language Technology

Language Resource Sharing is well established in the Language Technology research community, which relies on access to large quantities of linguistic corpora to advance work in this field. Here, concerted efforts have been made to amass resources in repositories, e.g. Linguistic Data Consortium (http://www.ldc.upenn.edu/) and OPUS (http://opus.lingfil.uu.se/index.php), as well as to index what is available in different repositories, e.g. META-SHARE (http://www.meta-share.eu/) and the European Language Resource Association catalogue (http://catalog.elra.info/). However, the language resources available largely come from publically funded sources or are one off data dumps offered by industry – both of which raise problems of resource staleness and maintenance. Furthermore, the wide variety of data formats found on these repositories remains a problem (Ide & Romary 2007). Several proposals are emerging, many based on LMF data categories, for RDF-based solutions to sharing and interlinking language resources. These include the Linguistic Annotation Framework (Ide & Romary 2004), LexInfo for linguistic-ontology linking (Buitelaar P. et al 2009), Lemon to support ontology localisation (Declerck T. et al 2010) and SEMbySEM formulti lingual lexicons (Falk I. et al 2010).

While term-bases and translation-memories can be directly reused in the localisation process, they are becoming increasingly relevant here as sources of parallel text corpora for training Statistical Machine Translation (SMT) engines. The popularity of the MOSES open source SMT toolkit (Hoang et al 2007) has spurred increased interest in the data management of parallel text, both to access it in sufficient volumes and to ensure it is sufficiently targeted and cleaned to yield high quality SMT performance in different translation application domains. Researchers have developed desktop tools (Clarke J.H. et al. 2010; Koehn, P. 2010) to ease the training of SMT engines and online data processing workflows tools (Tiedemann &Weijnitz 2010; Toral et al 2011) are emerging from current FP7 projects that also include resource upload and simple sharing mechanisms. Equivalent commercial offering are quickly emerging, e.g. SmartMate from ALS (http://beta.smartmate.co/) and the Microsoft Translator Hub (http://hub.microsofttranslator.com).

However, the effectiveness of these SMT training

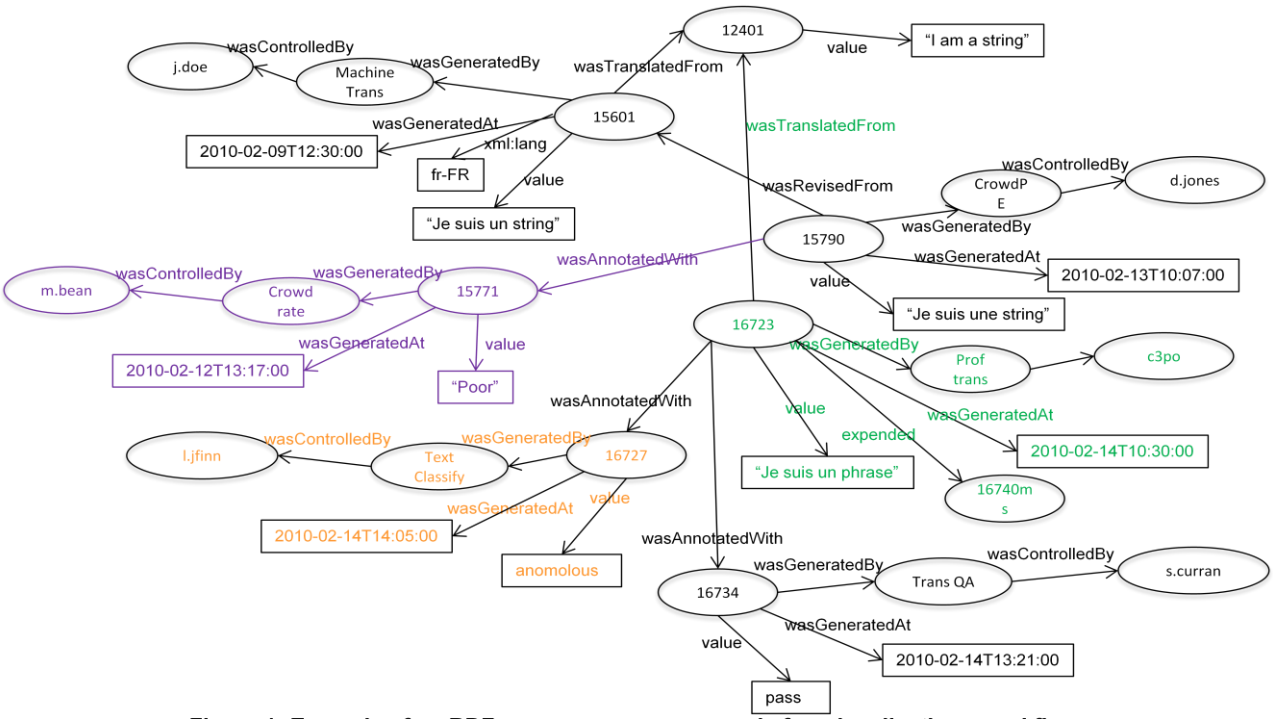

Figure 1: Example of an RDF open provenance records for a localisation nworkflow

technologies remains highly sensitive to the relationship of the training data to the localisation task at hand. The localisation industry is challenged across the board by the atomisation of content caused by the increasing popularity of smaller software products and shorter product lifecycles, e.g. for 'apps' distributed globally on smartphone and online social network platforms. Similar trends are visible in with user- and customer care-driven generation of product/platform support content (e.g. on wikis, user forums, and FAQ). This already erodes the effectiveness of term-base and translation memory reuse as the increased heterogeneity between jobs increases the corresponding domain, style and language mismatches that are observed between the available reusable resources and the incoming jobs. While SMT is more able to handle such mismatches than TM leverage, the quality of SMT output will similarly be eroded by a mismatch between training corpora and the content to be translated. Avoiding this requires low cost, but highly agile and responsive acquisition of training data. This can be achieved by continuous and targeted sharing of linguistic resources that are the by-products of localisation across the broadest range of LSPs and their clients. L3Data is therefore the means by which such systematic, fine-grained, pull-based sharing can be delivered on a large scale.

Until now there has not been an effective model for rapid language data exchange with suitable provenance needed to address commercial quality and access control concerns. The quality and relevance of such aligned text is increasingly being seen as a predictor than training data volume of the resulting SMT quality (Banerjee et al 2011; Vasiljevs 2010). L3Data sharing may therefore empower users to annotate aligned text in TMs with source and translation QA meta-data that will enable filtering of low-quality pairs. In addition, domain classification, part-of-speech tagging and linking to multilingual glossary entries may enable assembly of domain-targeted training data and supports the potential inclusion of lexical features in SMT training. Such would need to be supported by set of usage licenses (http://www.meta-net.eu/meta-share/meta-share/licenses) based on creative commons but tailored to the needs of language resource sharing.

5. Use Case

To provide some initial insight into how an L3Data model might be structured, consider the translation of a single sting, possible in the context of a small enterprise, localsing some web content, but wishing to minimize the cost involved. Figure 1 shows the RDF graph that may be build up using the open provenance model as this string passes through the following processes:

1. Being machine translated (node 15601)
2. The machine translation of the string being crowd-sourced posted edited resulting in a revised string (node 15709)
3. Further crowd-sourced effort is then used the annotate the translated string with a translation rating (node 15771)
4. Given that the crowd-sourced post-editing of the machine translation produced such poor results, it is decided in the end to opt for a professional human translation (node 16723)
5. A text analytics service is then used to compare the style of the translated string to a corpora in the desired style (node 16727)
6. Finally, because of the anomalous rating of the automate target language QA, a human QA is conducted, which is the end passes the string (node 16734).

Each of these nodes is an entity resulting from a process. Each provides the hub of a provenance record that details the process being performed (linked via attribute wasGeneratedBy), the agent performing the process (wasControlledBy) and the time at which is was completed (wasGeneratedAt). Each node can also be associated with values, for example the resulting string from a translation process (whether it be MT, crowd-sourced post-editing or professional translation), a specific attribute of the process, e,g. the time expended by the professional translation, or the value of an annotation (e.g. crowd-sourced, automated or professional QA assessments). The type ranges of these values will depend on the type of the entity. Relationships between the entities, e.g. wasAnnotatedWith, wasTranslatedFrom, can also be maintained. Overall, all the nodes use the same open provenance attributes, so SPARQL queries can later easily be formed, for example to extract parallel-text, but only with a quality pass status from a professional translation QA reviewer. Equally, queries could be formed to track progress of any number of segments through this process and to capture the path each took.

More significantly still, each node could be stored on a separate RDF triple store operated by different organization using the same schema as when it was stored on a single store. For example, the crowd-sourcing community may wish to keep ownership of its output, e.g. for gathering parallel text for training their own MT engine, while also making it available for an external clients. Equally, the professional translator may keep make the output of contracted translations and corresponding links to the source nodes available as RDF to authenticated clients, but may opt to put stronger access control on the processes performance data, e.g. the time it took to produce the translation, which may be kept confidential for competitive or pricing reasons.

This provides the ability to rapidly reconfigure language processing workflow while retaining the capability to maintain fine grained process monitoring and language resources sharing, due the ability, as linked data, to maintain links between individual RDF nodes.

6. Discussion

Efforts to create shared 'pools' of translation resources for industrial localization, such as that of the TAUS Data Association, have met with enthusiastic engagement from a variety of stakeholders. However, there have been significant challenges in the effective leveraging of the data shared. The narrow scope of content translated by most LSPs means finding a match for reusable resources with similar language, domain and style outside of a single client is rare. At the same time, centralised solutions raise doubts in potential participants about the motivation and sustainability of the approach (e.g. Google's June'11 announcement that it was withdrawing its free translation API) and down-stream copyright infringement of shared resources. L3Data addresses these concerns by opening up the technology for sharing at a fine-grained data level and by offering the potential to transparently expose the provenance of data reuse within a resource sharing network or via SMT training. This essentially places control of the sharing process in the hands of those contributing and benefiting from that sharing. This may in turn engender confidence in the process at a management cost accessible to even the smallest LSP, thereby overcoming the barriers raised by industry fragmentation and the burgeoning population of SME which make over 99% of LSPs. The resulting efficiencies in the translator productivity and the translation workflow will enable the industry to handle increased volumes of business, while offering clients more transparently assured quality in content translations.

L3Data sharing offers opportunities to forge novel business partnership between players in the localisation industry as well as with other adjacent markets. The primary opportunities will be:

Cooperating Networks of Language Service Providers and Translators: The share-alike model facilitated by L3Data allows Localisers to discover data which matches their specific requirements, if available, or to publish data they generate for others to find, reuse and enhance. This gives smaller LSPs the capacity to create networks of bi-directional sharing channels that provide them with access to searchable, online language resources on a previously unaffordable scale, thereby enabling them to be agile and cost-effective in competing for long tail localisation jobs. There is great potential for small LSPs to cooperate in bidding for work, possibly pooling language competencies and sharing data to bid for jobs of otherwise unattainable volume and language range. Individual translators will be able to pool resources in common domains and language pairs. Linked meta-data on the legal and copyright status of such data enables the auditing of the sharing of language resources between industry stakeholders at low cost and with a clear view of licenses.

Language Resource sharing for existing Value Networks: Localisation often occurs in existing value networks. Examples include: (a) network of platform technology vendors, e.g. MS Office, Apple's IOS, Google Android, and the developers of application based on these platforms; (b) enterprises and public sectors bodies working within a common regulatory framework, e.g. in the medical, legal or transport sectors and in the voluntary sector, e.g. translating open source software. These existing networks will benefit from terminology sharing and its use for domain specific SMT training.

Language Resource Curators: Localisation data, such as aligned text and term-banks typically requires substantial curation through data collection, cleaning, and value-add annotation before being suitable for reuse. If we consider the growth of L3Data sharing networks, such curation may become a viable commercial support service. An open data format for the core linguistic resources (i.e. parallel corpora and terminology) enables development of effective, multi-source linked data querying tools that can perform tasks such as cleaning, anonymising, quality rating, domain annotation and lexical annotation. The decentralised nature of linked data means that individual curators can add value to language data through links to meta-data that they generate and control access to. This offers the basis for controlled, but flexible commercial access to that meta-data, therefore offering new

commercial niches servicing the broader uptake in L3Data sharing. Interlinking itself, e.g. between aligned text and public term-bases, also becomes a viable value-add service, with the ease of interlinking triple stores deployed on the cloud.

Localisation System Integration: While free and open sharing of all localisation data on the web of data may require considerable medium-term repositioning within the industry, there is considerable near-term potential to provide platforms for consortia or value networks of LSPs and translators (commercial, public or voluntary) with shared interests in specific content domains. A key difficulty to sharing in such collaborative networks is the complex spectrum of different XML-based standards for exchange and their varying levels of tool compliance and performance. A key advantage of the Linked Data approach is the use of an open-world RDF model that allows systems to only consider the meta-data that they need, while safely ignoring any surplus information. This can considerably lower the cost of integrating new language technology services into such platforms, as well as of integrating with the client's content, customer and knowledge management systems.

Integration with knowledge management: Semantic Web and linked data technology is being applied with some success to improve the internal knowledge management of large and content intensive enterprises. This is primarily conducted by annotating content with an evolving ontology that represents and helps to organise and explore the complex knowledge structures used by the enterprise. This has a strong potential synergy with the L3Data approach. Already, for example, Interverbum's terminology management supports the construction of taxonomies. This could easily be enhanced to support ontologies with rich logical associations to support complex queries and mining of large volumes of content, regardless of the language involved.

7. Future Work

Building on our initial implementation of an L3Data platform (Lewis et al 2012) we aim to develop open interoperability with existing content management and localisation tools to provide commercially viable acquisition, pooling, preparation, interlinking, controlled sharing and audited reuse of L3Data as part of commercial localisation workflows. Seamless interoperation between content management systems and localisation tools remains a major challenge however. Though our implementation shows how an RDF provenance model can track and interlink content processes from both types of systems, their respective industry sectors still do not share many common semantics that can span this gap. One approach to addressing this is being undertaken by the recently formed MultilingualWeb-Language Technology working group at the W3C (Filip et al 2012). Building on the approach of the ITS standard, this group is developing a set of data categories addressing content meta-data as it rounds trips between content management systems, localisation workflows and language technologies such as SMT and text analytics. These semantics may then be usefully used to classify type of entities and processes used in L3Data province models. Indeed we see such as approach being more widely applicable to the end-to-end content value chain. Within the Centre for next Generation Localisation we have been developing broader semantic models of content processing that encompasses adaptive and content delivery, speech process and multimodal content interaction (Jones et al 2011). The L3Data sharing approach is well suited to rapid experimentation with novel forms of value networks that can exploit shared L3Data, including: crowd-sourcing of translations, annotation and quality assurance; translator cooperatives and skills banks; flash-team formation for on-demand SMT training; and both linguistic annotation and interlink maintenance-as-a-service. Combined with a broader view of content processing, and accompanying semantic models, such innovations will extend the impact of language resource sharing beyond the confines of existing localisation processes and into broad multi-lingual web content management.

8. Conclusion

To summarise, the proposed L3Data sharing aims to maintain links between fine-grained data resulting from the localisation process. These data inter-linkages are exposed as IRI/URLs to reap rewards of immediacy and consistency that are difficult to achieve when data interoperability relies on push-based file import and export functions between monolithic tools. Instead, as content is presented for localisation as a stream of new elements, consistency and quality maintenance for terminology and parallel text must become a continuous and mandatory rather than a periodic and optional activity. Such data and link maintenance is supported by linguistic technology services which can plug-in seamlessly to tools in the localisation workflow by matching their semantic service signatures with the semantics of the inter-linked content being processed.

9. Acknowledgements

This research is partially 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.

10. References

Abel, F., De Coi, J.L., Henze1, N., Koesling, A.W., Krause1, D., Olmedilla, D. (2007) Enabling Advanced and Context-Dependent Access Control in RDF Stores, International Semantic Web Conference 2007, LNCS 4825, pp. 1–14, 2007.

Alexander K., Hausenblas M. (2009) Describing Linked Datasets On the Design and Usage of voiD, the "Vocabulary Of Interlinked Datasets, Linked Data on the Web (LDOW2009), Apr 2009

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. (2007) Dbpedia: A nucleus for a web of open data, Journal of the Semantic Web LNCS 4825

Banerjee P., Naskar, S.K., Roturier, J., Way, A., van Genabith, J. (2011) Domain Adaptation in Statistical Machine Translation of User-Forum Data using

Component-Level Mixture Modelling. In Proceedings of the Thirteenth Machine Translation Summit

Bizer, C. Jentzsch, A. Cyganiak, R., (2011) State of the LOD Cloud, v0.3, Sept 2011, http://www4.wiwiss.fu-berlin.de/lodcloud/state/

Buitelaar P., Cimiano, P., Haase, P., Sintek, M.(2009) Towards Linguistically Grounded Ontologies, European Semantic Web Conference, LNCS 5554, pp 111-125

Clarke J.H., Weese, J., Ahn, B.G., Zollmann, A., Gao, Q.,Heafield, K., Lavie A.,(2010) The machine translation toolpack for loonybin: Automated management of experimental machine translation hyperwork, The Prague Bulletin of Mathematical Linguistics, 93:117–126, January 2010

Cruz-Lara S., Bellalem, N., Ducret, J., Kramer, I. (2004) Standardising the Management and the Representation of Multilingual Data : the Multi Lingual Information Framework. In Topics in Language Resources for Translation and Localisation. Elia Yuste (ed). John Benjamins Publishers, pp151–172

Declerck T., Buitelaar P., Wunner, T., McCrae, J., Montiel-Ponsoda, E., de Cea, A.G. (2010) lemon: An Ontology-Lexicon model for the Multilingual Semantic Web, W3C Workshop: The Multilingual Web - Where Are We? Oct 2010

Falk, I., Cruz-Lara, S., Bellalem, N., Osswald, T., Herrmann, V. (2010) Multilingual Lexical Support for the SEMbySEM project, LREC Workshop Language Resource and Technology Standards, pp19-329, 2010

Filip, D., Lewis, D., Sasaki, F., (2012) The Multilignual Web, WWW Confernece April 2012, Lyon, France

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF). LREC, Genoa

Ide, N. Romary, L., (2007) Towards International Standards for Language Resources, Evaluation of Text and Speech Systems, Springer 2007, pp 263-284

Ide, N. Romary, L., (2004) International standard for a linguistic annotation framework, Journal Natural Language Engineering, Vol 10 Iss 3-4, September 2004

Karamanis, N. Luz, S. Doherty, G. (2010)Translation practice in the workplace and machine translation, Conference of the European Association for Machine Translation (EAMT 2010), May 2010

Hoang, H., Birch, A., Callison-burch , C., Zens, R., Constantin, A., Federico, M., Bertoldi, N., Dyer, C., Cowan, B.,Shen, W., Moran, C., Bojar O., (2007) Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007

Jones, D., O'Connor, A., Abgaz, Y. M., & Lewis, D. (2011). A semantic model for integrated content management, localisation and language technology processing. In 2nd Workshop on the Multilingual Semantic Web. Bonn, Germany.

Koehn, P. (2010) An Experimental Management System, In The Prague Bulletin of Mathematical Linguistics, Sept 2010

Lewis, D., Curran, S., Doherty, G., Feeney, K., Karamanis, N., Luz, S., McAuley, J. (2009) Supporting Flexibility and Awareness in Localisation Workflows, Localisation Focus The International Journal of Localisation, 8, (1), p29 – 38

Lewis, D. O'Connor, A. Molines, S. Finn, L. Jones, D. Curran, S. Lawless, S. (2012) Linking Localisation and Language Resources, D. proc of Workshop "Linked Data in Linguistics", March 7 – 9, 2012, Frankfurt/Main, Germany

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

Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers J., Paulson, P. (2008) The Open Provenance Model: An Overview International Provenance and Annotation Workshop (IPAW), LNCS 5272, pp. 323–326, 2008

Muhleisen, H. Kost, M. Freytag, J. (2010) SWRL-based Access Policies for Linked Data, Access 576

TermBase eXchange (TBX), Systems to manage terminology, knowledge and content , ISO 30042:2008

Tiedemann, J. Weijnitz, P. (2010) Let's MT! — A Platform for Sharing SMT Training Data, In: Proceedings of the Third Swedish Language Technology Conference (SLTC-2010), pp 49-50

Toral, A. Pecina, P. Way, A. Poch, M.(2011) Towards a User-Friendly Webservice Architecture for Statistical Machine Translation in the PANACEA Project, Proc European Association for Machine Translation, May 2011, pp63-70

TMX (2005) 1.4b Specification OSCAR Recommendation, Localisation Industry Standards Association, 26 April 2005

Vasiljevs, A. (2010) Let's MT! — Platform for Online Sharing of Training Data and Building User Tailored Machine Translation, In: The Baltic Perspective, Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, vol. 219, October 7–8, 2010

Villata S., Delaforge, N., Gandon, F., Gyrard, A. (2011) An Access Control Model for Linked Data, , Workshop on Semantic Web &Web Semantics SWWS 2011

Wright, S.E. Budin, G. (1997) Handbook of terminology management, John Bejamins, 1997

XLIFF (2007) A white paper on version 1.2 of the XML Localisation Interchange File Format (XLIFF), Revision: 1.0, 17 Oct 2007