Open Data Vocabularies for Assigning Usage Rights to Translation Memories

David Lewis[1], Kaniz Fatema, Alfredo Maldonado, Brian Walshe, Arturo Calvo

[1]ADAPT Centre at Trinity College Dublin, Ireland,

E-mail: dave.lewis@cs.tcd.ie,

Abstract

An assessment of the intellectual property requirements for data used in machine-aided translation is provided based on a recent EC-funded legal review. This is compared against the capabilities offered by current linked open data standards from the W3C for publishing and sharing translation memories from translation projects, and proposals for adequately addressing the intellectual property needs of stakeholders in translation projects using open data vocabularies are suggested.

Keywords: Linked Data, Localization, Interoperability Standards

1. Introduction

To successfully curate data resulting from translation projects so that they can be more widely exchanged and leveraged beyond that project, the issue of managing the ownership of different aspects of this data and controlling the uses to which they can be put must be addressed. The intellectual property rights associated with translations are complex and potentially impacted by a number of different national and international laws and treaties. The issue has grown in prominence as the use of translation memories (TMs) has become widespread, in particular as fuzzy match scores against client-provided TMs has become a common discounting mechanism in pricing translation projects. The ownership of the TM is sometimes specified in translation project contracts. This may be between clients with sensitivities about content leakage or mature TM asset management strategies and Language Service Providers, to which translation projects are typically subcontracted. TM transfer also occurs between LSP and freelancers or between large multi-language LSPs and small single-language LSPs. TMs resulting from work of salaried employees are typically assigned as the intellectual property of the employer as part of the employment contract. However, as many translators work as contractors there is some awareness in the professional translator community of the intellectual property that is relinquished when LSPs request that TMs are handed over along with the translated project (Johnson2006). In many such situations the ownership of TMs is not clearly defined. LSPs or individual translators sometime exploit this lack of clarity to store the TMs from projects they have translated for use in future projects, potentially even with different clients. The complex nature of translation value chains, e.g. involving outsourcing to multi-language LSPd and thence to single language LSPs who in turn use freelancers, contributes to this lack of clarity. TMs need to be passed along this value chain for it to work effectively, but this makes defining and monitoring the conditions under which TM are stored a complex and potentially expensive administrative task for the parties involved. Several factors may contribute to the lack of clarity of TM and other data asset rights in translation contracts. Clients translating content for publication on the web may have less concern about TM leakage (i.e. confidential content leaving their control due to translation outsourcing and being used elsewhere), at least not once the content has been posted. Smaller translation clients and LSPs often do not have the management resources and expertise to assess the value of TMs and engage in negotiations on their value and protection in a contract. Further, LSPs and translators may not wish to draw client's attention to how the TMs may be used after a project finishes. As a result, the complexities involved coupled with the perception that the value of TM leverage may be marginal outside projects from the same client, means that there is little consensus on how TM ownership should be treated in contract negotiations, leaving the ownership of TM often unclear.

This lack of clarity is however a major impediment to the sharing of translation memory data. Consumers of such data will be wary of the risks of using translation memory if the ownership is unclear and the terms under which different uses of the data that can be undertaken is not well defined. Producers of data in the translation value chain may be reluctant to publish or exchange data for specific uses if their rights to do so are unclear to them.

One recent popular use of translation memories has been as a source of parallel text for statistical training machine translation (MT) engines. The use of such machine translation in translation projects has the potential to widen the opportunity for effective leverage data from a specific translation memory in translating content from different clients or domains. Recent efforts to share translation memories, such as the TAUS Data Association[1] and the LetMT! Corpora Repository[2] have primarily been driven by this motivation. While the usage rights for parallel text in such repositories are defined in terms and conditions, this is due largely to the centralised nature of these efforts that allows the rights to be more readily homogenised, via a common IP agreement in the case of TAUS and by selecting only corpora with a fully public license in the case of LetsMT!. Such homogenisation of usage rights is more

[1] https://www.taus.net/data/taus-data-cloud

[2] https://www.letsmt.eu/Login.aspx

challenging to achieve in situations where the publication of parallel text is decentralised by parties without a priori agreements such that variation more likely in the conditions under which the data may be shared and reused. Decentralised publication, however, has been shown by the linked open data community to support massive scaling in data exchange (Bizer et al 2009). However usage rights need to be declared in a way that can be readily indexed and searched alongside other data set meta-data. This enables parties seeking TM data to quickly determine if specific data is available for the use they need on terms that are agreeable. To this end we have proposed a simple integration of standard open data vocabularies for capturing parallel text resulting from a translation project, including meta-data on its provenance and usage rights.

To assess the adequacy of this approach for representing TMs for sharing, we analyse a recent report commissioned by the European Commission on Translation and Intellectual Property Rights (Troussel & Debussche 2014). In the context of the growing importance of machine-aided translation in the form of TM lookup and MT, this report examines the legal protection that can be extended to relevant data. This covers: the source documents subject to translation; the translations of those source documents; and the translation databases of sentence–aligned translations arising from the translation process. The report highlights the complexity of this issue, including the potential inconsistencies between different international and national legislation and treaties. It also highlights the importance of clarifying the ownership and usage rights of the elements of a translation database in a translation contract for the project that produces it.

Below we summarise relevant rights that should be addressed in a translation contract. While we do not aim to provide legal advice on translation contracts we do propose and analyse the form of machine-readable declarations of usage rights associated with relevant data resources at different points in the lifecycle of data used in translation projects.

2. Resource Rights related to Machine Aided Translation

Figure 1 outlines the primary data resources related to the use of a translation database in a translation project. It distinguishes between the data resources that are made available from previous projects commission by the client, resources from other third party sources and resources generated and used in the current project. Note there are other resources that can be considered, such as terminology resources and quality or productivity data, but these will be reviewed separately.

The relevant intellectual property rights identified by Troussel & Debussche (2014) are below summarised against this set of data resources. This summary aims to reflect the main conclusions of Troussel & Debussche in order to assess the sufficiency of the data schema used in addressing the most likely issues.

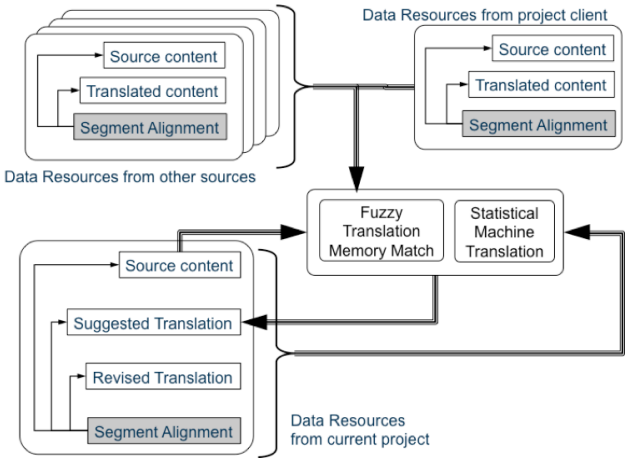

Figure 1: Use of different type of translation data in machine aided translation projects

The analysis does not aim to address the many national variations that the report identifies (which is by its own admission incompletely), nor the exceptions available for specific content domains, such as scientific works or public sector information documents. Where relevant and in the absence of relevant translation contract terms, the property rights are ascribed to the workers involved, though it is assumed that these rights fall to their employers where they are salaried employees. In reference to the data resources identified in figure 1 the usage right issues are as follows.

Data Resources from the project client or from other sources:

This consists of:

Source Content: The copyright belongs to the authors of that content and grants them economic rights over the: reproduction, adaptation, alteration, distribution, communication to the public, use in derivative works (which may include databases) and most importantly over translation. The authors also may hold moral rights, whereby their good character may be protected if harmed through treatment of the content they produce. Of relevance here may be a misleading translations that damages the integrity of the work and thereby the reputation of the author.

Translated Content: The translator owns the copyright over the translated content. As the right to authorise translation is held by the holder of the source content's copyright, the translator may be in breach of this right if permission to translate is not explicitly granted. This however may not necessarily prejudice the copyright over the translation held by the translator.

Segment Alignment: An outcome of the translation process is the alignment of source and target language segments, which can be captured in a database forming a translation memory or parallel text data resource. While this database is not in itself a subject of copyright, in the EU it may be protected under the EU Database Directive. This offers in one part copyright protection over the design of the database schema. However, as the structure of translation memories and the process of segmentation are widely understood and even subject to standardisation, e.g. the Translation Memory eXchange

XML format, there seems little opportunity to show originality in the schema design of TMs. However the Database Directive also support a Sui Generis protection over the database that is granted if its creator can demonstrate substantial investment made in obtaining, verifying or presenting the content of the database. Sui Generis protection grants rights over the extraction of substantial portion of the database (similar to copyright for reproduction) and reutilisation (similar to copyright for communication to the public). Given the widening use of translation management tools provided in the cloud by LSPs and the curation and quality assurance effort undertaken by LSPs in assembling translation memory, the breakdown of effort in assembling a TM database (as opposed to generating the translation) would typically be weighted towards the LSP, giving them possibly a stronger claim than translators to Sui Generis rights over the alignment.

Data Resources from Current Project

Source Content has rights following the same rules as for content from other projects.

Suggested Translations generated by machine aided translation services are not subject to any copyright protection as they are automatically generated rather than the result of creative human effort. The generation of suggested translations for individual project source segments using either TM fuzzy matches or statistical MT requires TM or parallel text as the enabling data resource. This therefore requires the provider of the machine-aided translation service to obtain the rights to use the data. This involves being granted rights by the source and the translation copyright holders and potentially the Sui Generis rights holder of the translation memory database.

Revised Translations provided by a human post-editor could be subject to copyright protection as for translated content in general. However, if the machine translated output is of a high quality, such that very little post-editing is required, then the claim of the post-editor to providing creative input to the generation of translation may be challenged, weakening the claim to copyright protection over the revised translation.

Segment Alignment if the project includes the sentence-by-sentence alignment between source content, machine suggested translation and revised translation resulting from post-editing. There are several uses of this data resource that need to be considered in supporting IP declaration and assignments:

- Once the project has been completed, the segment alignment between the Source Content and the Revised Translation would act as a data resource from a prior project that can then be reused in subsequent projects. This reuse would require assigning rights to reuse from the source author, the translator/post-editor and the TM database rights holder.
- The same assignment of rights would apply if the intention of the client or the LSP was to communicate the TM as a database to the public. In web site translation, it may be the case that both the source and the target content are communicated to the public with some conditions, e.g. for attribution. It is unclear however how separate copyright over published source and target documents may impact the use of that content by third parties to generate parallel text using sentence splitting and alignment software to recreate their own version of the aligned text. Publishing the segment alignment alongside the source and target text would in these circumstances allow the LSP to assert rights and specify assignment conditions over the translation memory that would otherwise remain latent but mine-able in the published content. Clear declaration of these rights published alongside the content would encourage those seeking the parallel data to abide by the conditions of use associated with the aligned data.
- The alignment between the Source Content and the MT-generated Suggested Translations is of value in assessing and tuning the performance of the MT engine, for example in relation to catching issues with out of vocabulary words or specific terminology translation. As this is a different use compared to the use of source and aligned revised translation segments for MT training or TM databases, it may need different usage rights associated with it.
- The segment alignment of the MT-generated Suggested Translation and human-generated Revised Translation allows the edit distance between the two to be measured. This can be a valuable data asset in predicting the effort expended by post-editors in turning the machine-aided translation into something they consider of acceptable quality for a translation project. This data could be supplemented with other operational data such as the time taken to post-edit the segment, the keystrokes involved and any translation quality review annotation to gain a more detailed picture of the efficacy of the suggested translation in reducing the time and effort required to post-edit it. If collected systematically, such data could also help justify claims for Sui Generis rights over the alignment and copyright over the translation but documenting the human effort and creative acts involved.
- With cloud-based translation management systems or computer assisted translation tools , it is already common practice to reintegrate Revised Translations into a project TM database so it can be leveraged subsequently in the project. Similarly, some tools are exploring the similar use of segment post-edits for the iterative retraining of the MT engine during the translation project. From an IP point of view this requires the same in-project reuse of Revised Translation segments. Therefore, the agreement of the source author (typically via the project client), the post-editors producing the Revised Translations and the live TM database owner needs to be secured for this purpose. This IP assignment must accommodate several post-editors working on the project, and that the MT provider may be a separate organisation to the LSP that is assembling the

project's TM database.

Previously we had suggested using linked open data vocabularies to capture data from various stages of a translation workflow in order to facilitate its reuse (Lewis et al 2012). We have defined an Linked Language and Localisation Data (L3Data) Schema (Lewis et al 2015) that uses open data schema from different standardisation activities related to the W3C Data Activity. Of those, the schema relevant to this analysis are as follows, with the schema name space use to identify specific properties given in parenthesis:

- Data Catalogue Vocabulary (dcat) (Maali & Erickson 2014) which is a W3C Recommendation for cataloguing open data sets.
- Dublin Core (dcterm) (Powel et al 2007), which is well established meta-data for documents and is referenced from DCAT.
- Open Digital Rights Language (odrl), (McRoberts & Rodrigues Doncel 2015) which aims to support machine readable licensing terms – currently a W3C community group document. ODRL in turn, following best practice in open data vocabulary design, makes use of concepts from other schema, of relevance here is the Creative Commons (cc) vocabulary available at creativecommons.org/ns.

ODRL Action and definition	Use in TM and MT asset management
odrl:use The Assigner permits/prohibits the Assignee to use the Asset as agreed. More details may be defined in the applicable agreements or under applicable commercial laws. Refined types of actions can be expressed by the narrower actions.	A general action capturing the widest range of uses for which rights can be assigned. Does not offer the specificity needed in some TM value chains (see discussion)
odrl:grantUse The Assigner permits/prohibits the Assignee to grant the use the Asset to third parties. This action enables the Assignee to create policies for the use of the Asset for third parties. nextPolicy is recommended to be agreed with the third party. Use of temporal constraints is recommended.	Important for the control of the broad action of 'use' when reassigned along a value chain, e.g. a client can grant 'use' to an LSP, but engage in the definition of the terms under which the 'use' can be passed onto contract translators working on the project. The secondary license is defined in a odrl:Policy reference by the odrl:nextPolicy action (see below).
odrl:compensate The Assigner requires that the Assignees compensates the Assigner (or other specified compensation Party) by some amount of value, if defined, for use of the Asset.	Potentially useful for controlling the project price discounting terms for an LSP using a client's TM.
odrl:acceptTracking The Assigner requires that the Assignees accepts that the use of the Asset may be tracked. The collected information may be tracked by the Assigner, or may link to a Party with the role function "trackingParty".	This can used by a client to require an LSP to track the use of a TM by subcontractors. It could also be use to specify that the use of translated segments in training different MT engines be tracked and reported. Similarly it may allow the use of revised translations may be tracked, e.g. if posted as content on a public web site, the terms and conditions specify that web analytics and possible A/B testing may be employed in the assessment of translation quality.
odrl:aggregate The Assigner permits/prohibits the Assignees to use the Asset or parts of it as part of a composite collection.	TMs are often combined when used for MT training, so this practice can be controlled using policies for this action.
odrl:annotate The Assigner permits/prohibits the Assignees to add explanatory notations/commentaries to the Asset without modifying the Asset in any other way.	Could be used to control the use of quality annotations of the translated segments, e.g. using a open quality framework such as the Multidimensional Quality Metrics framework[3] or terminological annotations of source or translated segments.
odrl:anonymize The Assigner permits/prohibits the Assignees to anonymize all or parts of the Asset. For example, to remove identifying particulars for statistical or for other comparable purposes, or to use the asset without stating the author/source.	It is common practice for sets of translated segments to be recorded with meta-data on the identity of the translators who produced them. This personal identification meta-data is both commercially sensitive in the translator subcontracting market, and could also contravene workplace agreements, and needs to be controlled. This action allows control over the exchange of such personal meta-data with translation data.
odrl:archive The Assigner permits/prohibits the Assignees to store the Asset (in a non-transient form). Constraints may be used for temporal conditions.	Could be used to control the period for which a TM can be stored regardless of the use to which it is put. Can help control the long term storage of such data resource in situations where future uses are difficult to predict.

[3] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

odrl:attribute The Assigner requires that the Assignees attributes the Asset to the Assigner or an attributed Party. May link to an Asset with the attribution information. May link to a Party with the role function "attributedParty".	This action enables assigner of rights to resources, such as TM, which they make publically available to stipulate that public acknowledgement of that use is made, and thereby reputational benefits accrue. This is a common clause in many public TM licenses.
odrl:copy The act of making an exact reproduction of the asset.	This could be used to control the ability to make a copy of a TM outside of a TMS, e.g. via a TMX export feature.
odrl:delete The Assigner requires that the Assignees permanently removes all copies of the Asset. Use a constraint to define under which conditions the Asset should be deleted.	This could be used to specify conditions under which TMs provided to an LSP should be deleted, e.g. on termination of a long running contract.
odrl:derive The Assigner permits/prohibits the Assignees to create a new derivative Asset from this Asset and to edit or modify the derivative. A new asset is created and may have significant overlaps with the original Asset. (Note that the notion of whether or not the change is significant enough to qualify as a new asset is subjective).	This could be used to control some common transforms conducted on TMs, e.g. removing mark-up or decapitalisation prior to use in MT training.
odrl:distribute The Assigner permits/prohibits the Assignees to distribute the Asset.	Could be used to control distribution to third parties (e.g. constrained to classes such as academic parties) and public communication of assets such as TMs.
odrl:ensureExclusivity The Assignee requires that the Assigners ensure that the permission on the Asset is exclusive to the Assignee.	This could be used to ensure that an LSP assigned a TM does not, for example, use that TM to benefit other client's projects or pass to other collaborating LSPs.
odrl:extract The Assigner permits/prohibits the Assignees to extract parts of the Asset and to use it as a new Asset. A new asset is created and may have very little in common with the original Asset. (Note that the notion of whether or not the change is significant enough to qualify as a new asset is subjective).	This could be used to control extraction of cross-lingual terminology or phrase bi-text from a TM.
odrl:give The Assigner permits/prohibits the Assignees to transfer the ownership of the Asset to a third party without compensation and while deleting the original asset.	Could be used to control the non commercial distribution of TMs to third parties.
ordl:index The Assigner permits/prohibits the Assignees to record the Asset in an index. For example, to include a link to the Asset in a search engine database.	Could be used to control the ability to use assigned TMs in TM lookup, concordancing or word alignment software, though more specific profiles may help.
odrl:inform The Assigner requires that the Assignees inform the Assigner or an informed Party that an action has been performed on or in relation to the Asset. May link to a Party with the role function "informedParty".	Allows control of the observation of the specific uses to which a TM asset is used. For example could be used to ascertain the risk of an LSP misusing a TM without having to rule these uses out in detail beforehand.
ordl:lease The act of making available the asset to a third-party for a fixed period of time with exchange of value.	A means of controlling the period of use of an asset, e.g. the use of the TM beyond the end of a project.
odrl:lend The act of making available the asset to a third-party for a fixed period of time without exchange of value.	Similar temporal control to using odrl:lease, but without the presumption of commercial or other value exchange.
odrl:modify The Assigner permits/prohibits the Assignees to update existing content of the Asset. A new asset is not created by this action. This action will modify an asset which is typically updated from time to time without creating a new asset like a database. If the result from modifying the asset should be a new asset the actions derive or extract should be used. (Note that the notion of whether or not the change is significant enough to qualify as a new asset is subjective).	Could be used to control the update of a client's TM database, e.g. in integrating revised translation from post-editing into a project TM or MT retraining process. It allows modification without relinquishing control over the asset.
odrl:nextPolicy The Assigner requires that the Assignees grants the specified Policy to a third party for their use of the Asset.	This allows the assigner to specify a policy under which an action can be assigned onwards to a third party, so important for allowing clients to control the terms under which TMs are assigned by LSPs to contract translators.
odrl:obtainConsent The Assigner requires that the Assignees obtains explicit consent from the Assigner or a consenting Party to perform the requested action in relation to the Asset. Used as a Duty to ensure that the Assigner or a Party is authorized to approve such actions	Could be used to control the actions permitted on assets assigned by a client or an LSP for actions where the consent of the specific translator or content author is required, e.g. in cases where they are not salaried employees and transfer of ownership was not established

on a case-by-case basis. May link to a Party with the role function "consentingParty".	in the work contract.
odrl:read The Assigner permits/prohibits the Assignees to obtain data from the Asset. For example, the ability to read a record from a database (the Asset).	Could be used to control general access to a TM, as part of restricting its use to specific TMS based functions such as TM look-up and concordancing.
odrl:reproduce The act of making an exact reproduction of the asset. The Assigner permits/prohibits the Assignees to make exact reproductions of the Asset.	Could control ancillary copies of TM being made from a TMS, especially when the TMS provides sufficient search and processing features to translators. It can therefore control the export of TMs by assignees into third party tools with unknown vulnerabilities.
ordl:reviewPolicy The Assigner requires that the Assignees have a person review the Policy applicable to the Asset. Used when human intervention is required to review the Policy. May link to an Asset which represents the full Policy information.	Useful to control the human workflow of checking licenses before performing specific actions of assigned assets.
odrl:secondaryUse The act of using the asset for a purpose other than the purpose it was intended for.	This could be used to restrain the use of TM, e.g. for TM lookup only, without having to specify other statistical leverage that could undertaken with a TM, e.g. not just SMT training, but multilingual terminology mining or monolingual content analysis.
odrl:sell The Assigner permits/prohibits the Assignees to transfer the ownership of the Asset to a third party with compensation and while deleting the original asset.	Allows control over the resale of TMs.
odrl:transfer The Assigner transfers/does not transfer the ownership in perpetuity to the Assignees.	Useful to control the permanent transfer of assets, e.g. of a client TMs to a TM aggregation service.
ordl:transform The Assigner permits/prohibits the Assignees to make a digital copy of the digital Asset in another digital format. Typically used to convert the Asset into a different format for consumption on/transfer to a third party system.	Could be used for controlling the transformation of source content, e.g. from HTML to XLIFF, and bi-text, e.g. from TMX to CSV.
odrl:translate The Assigner permits/prohibits the Assignees to translate the original natural language of an Asset into another natural language. A new derivative Asset is created by that action.	Offers important control over source segments to allow the assignee to translate it. The derivative asset are the translated segments.
cc:ShareAlike The act of distributing any derivative asset under the same terms as the original asset.	ShareAlike is a common model for many forms of open data exchange and so could be relevant for publication of TM, especially if resulting from crowd-source translations.ec

Table 1. ODRL actions and relevant use in resources relevant to translation

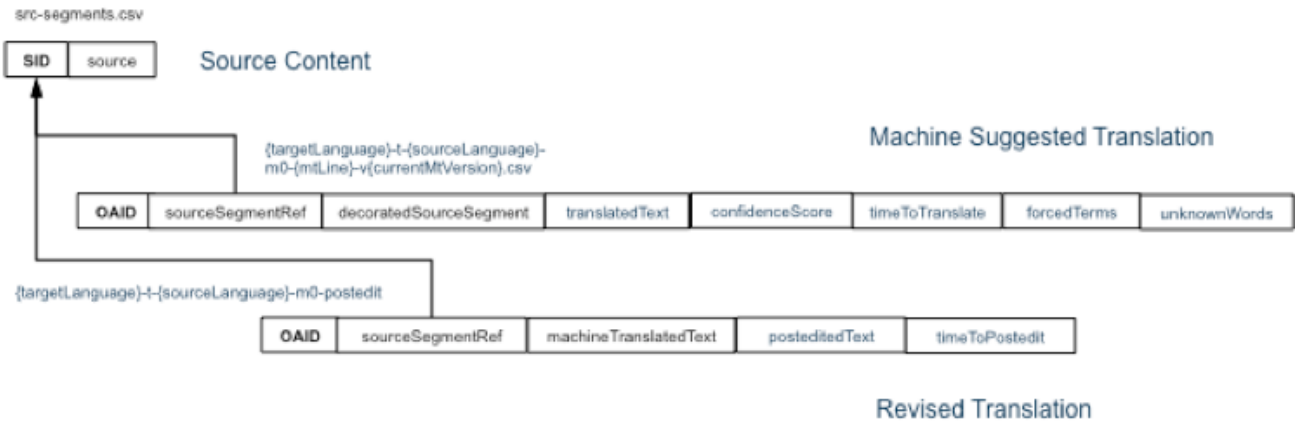

Figure 2. Sample L3Data CSV-on-the-Web Schema

The basic IP management mechanism is to store the relevant data in a CSV file with meta-data from the schema defined in the W3C CSV-on-the-Web meta-data specification (Pollock et all 2015). This is used to define the CSV file as the dcat:Distribution object, which is a downloadable datset as per the DCAT vocabulary. This object can have a dcterm:rights property that in turn can

point to an ODRL file. An ODRL file can define an odrl:Policy object with attributes defining a rule under which rights may be assigned. It can specify: an odrl:assigner attribute, identifying the entity granting the permission; an odrl:permission attribute, specifying the action for which this rule assigns rights over for the subject resource; and an odrl:prohibition attribution specifying the actions that are specifically not granted. Further, a odrl:Policy can define a odrl:Duty object indicating actions the assignee of a right must undertake in realising that right, e.g. providing payment or attributing the creator of the asset in any publication that uses it.

Key to the appropriate formulation of machine readable rights policy in ODRL therefore is the definition of the odrl:Action object defined in the vocabulary. Policy rules define obligation or constrains related to such actions. ODRL defines 61 instances of the odrl:Action object covering a range of activities over data and digital content which are deemed useful in assigning rights in a machine readable format. Of these, those listed in Table 1 are identified in as being relevant to the assigning right to assets corresponding to the data resources defined above that are relevant to translation projects. Table 1 presents the definition for each of the actions and an explanation of how it could be used in the context of translation project data.

Figure 2 provides the schema used for an L3Data scenario tested in the FALCON Project. The source, machine suggested translations and revised translations are captured in CSV files with a row for each segment. The two translation tables are modelled as annotations of the corresponding rows in the source content table, containing the translations and other annotations related to the translation process for each segment. Each table is accompanied by a CSV-on-the-Web meta-data file, which identifies the table as a dcat:Distribution and gives the pointer to the relevant ORDL file.

3. Discussion

The combination of DCAT and ODRL as defined in the L3Data Schema to annotate data structured according the CSV-on-the-Web provides a comprehensive and flexible mechanism for assigning rights to data resources used in and resulting from translation workflows that leverage machine aided translation technology. The following issues are raised however by the above analysis:

Dataset granularity for rights annotation

The L3Data schema identifies an individual CSV file as a dcat:Distribution object, with a reference to a single specified ODRL file. In situations with multiple CSV files have the same distribution rights this is an efficient mechanism, since only one ODRL file needs to be managed. This is advantageous since the legal aspects of assuring the correct configuration of ODRL files may make their management and quality assurance an expensive task. However, this means it is complex to express differing right for different attributes in a CSV resource recorded in different columns. For example in the schema scenario shown in figure 2, the revised translation has columns for both the segment translation and operational meta-data such as the time it took to post edit each segment. It can easily be envisage that the translator undertaking the post-editing may wish to specify different usage conditions for the translation text and for the performance-related meta-data, in terms of how it could be used and to whom it could be released. DCAT does not offer a mechanism for nesting dcat:Distributions, so all parts of it, i.e. the entire CSV file, must have the same rights annotation. It is possible to differentiate odrl:Policy object over different parts of a CSV file by specifying that the policy applied to a specific odrl:target using a URL. Such a URL can reference an individual column using a column fragment identifier, e.g. for a revised translation CSV file en-t-fr-m0-postedit.csv (using the schema from figure 2) the translation data can be referenced as en-t-fr-m0-postedit.csv#col=posteditedText and the post-editing timing information can be referenced as en-t-fr-m0-postedit.csv#col=timeToPostedit. However, as the reference from the odrl:target attribute is a URL, one such declaration needs to be added for each CSV table which references that ODRL file, thereby complicating the management of the ODRL file. The recommended approach therefore is to separate out data that requires different license conditions into different CSV files. This will ensure any access control mechanism that process the ODRL file will be able to unambiguously determine when the entire CSV file should be accessed, while constraining the scale of the ODRL management and checking task.

Specific ODRL action definitions for reuse in TM leverage and MT training

The ODRL action primitives provide a set of actions that could be used to constrain the technical uses via which translation related resources can be leveraged. For instance, prohibiting indexing would effectively prevent the use of bi-text in TM lookup and MT training, since indexing is a fundamental part of these activities. Similarly, prohibiting extraction, may permit TM lookup but constrain use in MT training. However, these are highly technical mappings and therefore suffer from both being barriers to full understanding of their implications for translators and translation project managers. They may also be circumvented by innovation with technical techniques or arguments over the legal interpretation of technical terms. It is therefore recommended that:

- ODRL be used with domain specific action definitions documented as an ODRL Profile.
- The translation community establish some consensus on action primitives that can be used in ODRL policies that are relevant to its concerns. For example, primitives that allow policy rules to distinguish TM lookup from MT training could be likely candidates, offering some protection to translators while not being too complex to understand and interpret.

Assigning rights to terminology

The analysis in (Troussel & Debussche 2014) focuses on translation memories and does not examine the case of multi-lingual terminology in detail. While terminology generation in translation projects is seen as a specialised task conducted by terminologists, some terminology management tools are now integrated with translation management tools so that the management of term lifecycle can be integrated with a project:. This can be useful in support of human translation and post-editing; in support of automated term extraction specific to a project and in support of guiding MT engines on term translations. Further, open vocabularies enable third party lexical-conceptual resources such as Babelnet to be leveraged to support the translation process. Support is provided in help post-editors understand possible definitions of newly identified terms and in accessing potential translations of new terms for the benefit of both post-editors and machine translation engines. While resources such as Babelnet (Navigli & Ponzetto 2012) can be commercially licensed, they depend on a large extent on the aggregation and processing of publically funded or crowd-sourced lexical, terminological or ontological resources. While this use may permissible in the terms under which these resources are published, the sustainability of these approaches may be damaged if the attribution for use of the source resources is lost as the original lexical-conceptual knowledge is aggregated and used via web services to answer specific queries or annotate text. This lack of attribution may disincentivise non-commercial producers of lexical-conceptual knowledge, especially in more specialised domains and less well resourced languages. Further, the validation, including negative validation of the use of terms and terms translation in specific segments of a translation project may be a source of valuable 'in-context' annotations for terms. This may be used to improve text analysis services that provide lexical-conceptual or terminological annotations to third parties, e.g. using term extraction, named entity recognition, entity linking, part of speech tagging and word sense disambiguation techniques. However, if the assertion and assignment of rights over these term-in-context annotations is not easily captured, then this inhibits attribution or compensation for use of this data between parties and disincentivises its capture as reusable datasets. The L3Data Schema provides for separate recording of such term-in-context validation data in a CSV annotation table, and thereby enabling the assertion of Sui Generis database protection rights over it by LSPs, translators or terminologists. However, this is not a well-established aspect of data exchange in the translation industry, and the value and pricing of text annotation services are still poorly understood. It is therefore recommended that further study be conducted into: the relevant value of term-in-context validation from translation projects in the improvement of text analysis software performance and the different uses of text analysis software in the translation industry and more widely in the multilingual web content access industry.

4. Acknowledgements

This work has been supported partially by the European Commission as part of the FALCON project (contract number 610879) the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

5. References

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

Lewis, D., Maldonando, A., Wlashe, B., Fatema, K., Calvo, A. (2015) Revised L2Data Federation Platform Release, FALCON Project deliverable D2.3, 30-6-2015, http://falcon-project.eu/wp-content/uploads/2015/11/D2.3-submit.pdf

Lewis, D., O'Connor, A., Zydroń, A., Sjögren, G., Choudhury R., (2012) On Using Linked Data for Language Resource Sharing in the Long Tail of the Localisation Market, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)

Johnson, A.E. (2006) What right does a translation agency have to demand delivery of the TM upon completion of a translation project?, Panel contribution to Proz.com Regional Conference, Edinburgh 10-11 November, 2006, retrieved 17/2/2016 from http://www.proz.com/edinburgh/AEJohnson_TMs_and_copyright.pdf

Maali, F., Erickson, J. (2014) Data Catalog Vocabulary (DCAT), W3C Recommendation 16 Jan 2014, https://www.w3.org/TR/vocab-dcat/

McRoberts, M., Rodriguez Doncel, V. (2015), ODRL Version 2.1 Ontology 5-3-2015, https://www.w3.org/ns/odrl/2/ODRL21

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

Pollock, R., Tennison, J., Kellog, G., Herman, I. (2015) Metadata Vocabulary for Tabular data, W3C Recommendation 17 Dec 2015, https://www.w3.org/TR/2015/REC-tabular-metadata-20151217/

Powel, A., Nilsson, M., Naeve, A., Johnston, P., Baker, T. (2007) DCMI Abstract Model, 4-6-2007, http://dublincore.org/documents/2007/06/04/abstract-model/

Troussel, J-C, T., Debussche J. (2014) Translation and Intellectual Property Rights (Report by Bird & Bird for the European Commission DG Translation). Luxembourg: European Commission.

doi:10.2782/72107