D4.2: LINGUISTIC TASK USABILITY AND REUSE EFFICACY RESULTS

Joss Moorkens, Ankit Srivastava, Balázs Benedek, David Lewis, Kaniz Fatema

Distribution: Public Report

Federated Active Linguistic data CuratiON (FALCON)

FP7-ICT-2013-SME-DCA

Project no: 610879

Document Information

Deliverable number:	D4.2
Deliverable title:	**Linguistic task usability and reuse efficacy results**
Dissemination level:	PU
Contractual date of delivery:	31st May 2015
Actual date of delivery:	17st September 2015
Author(s):	Joss Moorkens, Ankit Srivastava, Balázs Benedek, David Lewis, Kaniz Fatema
Participants:	DCU, SKAWA, TCD
Internal Reviewer:	TCD
Workpackage:	WP4
Task Responsible:	T4.2
Workpackage Leader:	Joss Moorkens

Revision History

Revision	**Date**	**Author**	**Organization**	**Description**
1	12/10/2014	David Lewis	TCD	Template
2	14/08/2015	Joss Moorkens	DCU	Trial evaluation documentation
3	05/09/2015	Ankit Srivastava	DCU	Automatic Evaluation documentation
4	6/09/2015	David Lewis, Kaniz Fatema	TCD	Security and intellectual property management assessment
5	16/9/2015	David Lewis	TCD	Review and corrections

Contents

1. Executive Summary

This document summarises the second round of evaluation conducted by the FALCON project. It reports on human translator evaluations using the portal made available for accessing the FALCON showcase system and automated assessment of the resulting translation quality against a reference. In addition, the document provides an assessment of the system's security configuration and the intellectual property management capability of the data standards used.

2. Introduction

This deliverable reports on human evaluations carried out using the FALCON portal, integrating the XTM Cloud Translation Management System, the Skawa' EasyLing website translation platform, and Interverbum's Termweb terminology tool, along with term extraction, purpose-built Statistical Machine Translation (SMT) engines that are retrained based users' post-edits, and segment reordering for SMT optimisation. The deliverable reports on an initial pilot study, testing the robustness of the tool integration at the core of FALCON. Thereafter, it provides quantitative and qualitative results of translator evaluations in three language pairs: English-German, English-Spanish, and English-French. The evaluation participants used the FALCON portal in several scenarios, in randomised order, to test the various elements of the FALCON portal. They post-edited previously translated public documents, so that both automated scores and task performance (e.g. segment translation time) could be captured and compared. In addition to measures of participant productivity, the performance of the SMT system was measured in terms of translation quality using standard automatic evaluation metrics, i.e. Translation Edit Rate [TER – Snover et al., 2006] and BLEU [Papineni et al., 2002].

The deliverable also provides a summary assessment the system level security of the platform, which is based on the security measures of the existing commercial platforms extended with a shared login feature. An assessment of the intellectual property requirements for data used in machine aided translation is provided based on a recent EC-funded legal review [Bird&Bird 2014] and this is compared against the capabilities offered by the open data standard proposed for publishing and sharing translation memories from translation projects.

3. Trial Evaluation Setup & Results

3.1. Pilot Study

Prior to the main evaluation, a pilot study was carried out in order to test the functionality of the integrated FALCON portal, to test the evaluation methodology, and to make adjustments and refinements as necessary. For the pilot study, one English to Spanish translator at the Skawa offices in Budapest post-edited a test set of 125 segments from the News-Test corpus[1] within the FALCON interface with the assistance of a Moses-based SMT engine trained purposely on different data from the DGT (Directorate General of Translation) corpus. This first test set was completed and the post-edited text was used to retrain the SMT engine. The translator then post-edited two smaller 50-segment test sets from the same News-Test corpus: one test set was post-edited to measure the perceived SMT quality improvement, and the other had been reordered to favour segments that were expected to improve the SMT quality, while using the Slimview feature to see the segments in context.

The pilot study revealed some issues with the initial proposed methodology. Participant speed was slower than expected, necessitating test set size to be minimised. Some technical issues within the portal required

[1] These are news articles from 2013 extracted from various online publications and made available at http://www.statmt.org/wmt15/translation-task.html for use in MT tasks.

debugging, such as communication problems between the XTM cloud interface and Slimview, and optimisation of SMT engines on the FALCON server. The participant required more information about the FALCON portal and the evaluation process than had been initially provided. As a result, further documentation was written and discussed with participants prior to beginning the full evaluation.

3.2. Evaluation Method

The data used for this evaluation were three test sets of 100 segments (TS1,2,3) and three test sets of 50 segments (TS4, 5, 6) from the News-Test corpus. The reason for the discrepancy in test set size is that 50 was considered an insufficient number of segments to effectively retrain an SMT engine, although our expectation was that retraining would only prove effective over a longer term than the few days covered in this study. The test sets were prepared for post-editing (PE) by changing the format to HTML (to publish for Slimview and to tag UTF encoding), and then tested for homogeneity using Wordsmith analysis. Wordsmith found that the mean word length for the source test sets (4.70 to 4.86 characters) was representative of the whole (4.75 characters), and similarly the mean number of words per segment (17.67 to 22.91) and the standardised type-to-token ratio (41.45 to 46.88) was consistent throughout the test sets and representative of the text as a whole. The SMT engines were trained on the DGT corpus (so that the test was out of domain). Language pairs used were English to French, English to German, and English to Spanish.

The evaluation performance was measured in several different ways. Using timestamp information that the XTM development team added to the post-translation XLIFF files, temporal productivity measurements were calculated per segment and per task to provide a comparative words-per-second measure. Segments taking more than 5 minutes were considered outliers and discounted. SMT engine performance after retraining was measured using the BLEU automatic evaluation metric, and technical post-editing effort was measured using the TER metric.

Post-editing productivity was tested in the following conditions:

1. **PE using FALCON interface as baseline**
2. **PE re-ordering the segments with *Slimview***
3. **PE re-ordering the segments without *Slimview***
4. **PE with terms from automatic source term capture**
5. **PE using MT improved through iterative retraining following minimal reordering**
6. **PE using MT improved through iterative retraining following reordering**

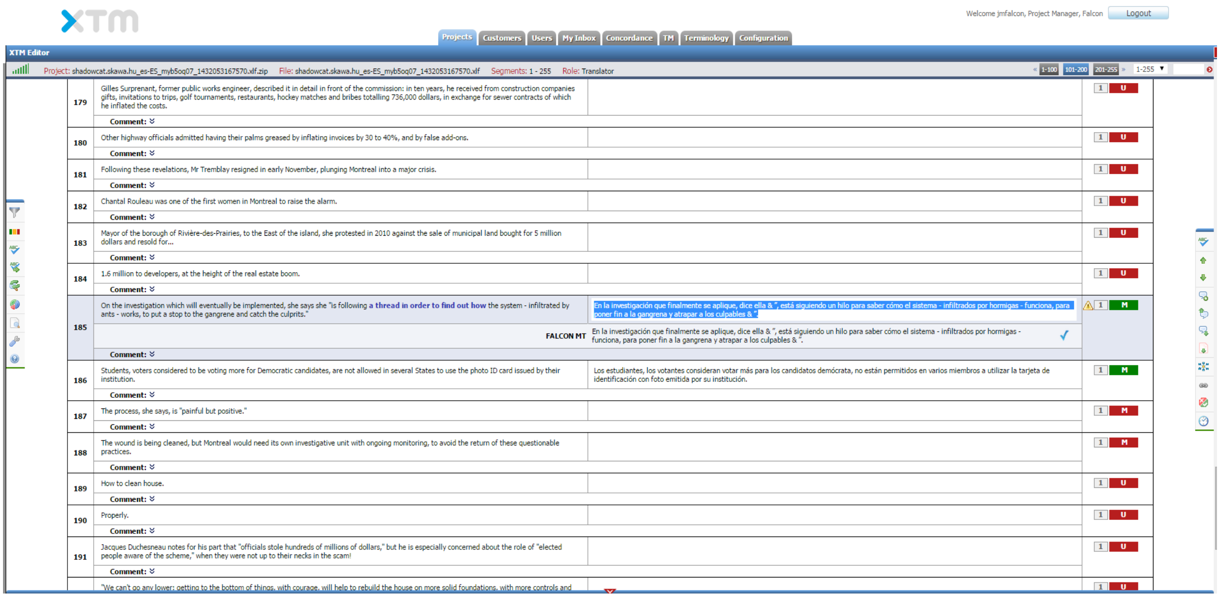
We engaged three participants per language pair, giving them detailed instructions for completing the tasks, and a plain language explanation of the study as stipulated by the Dublin City University Research Ethics Committee. Three of the participants dropped out of the study, and one more failed to complete all tasks within the allotted time, leaving five who completed all tasks as requested. Participants were asked to complete the tasks in the order specified in Table 1. The tasks were set up as numbered projects within the FALCON portal and assigned to the participants. Once the tasks were completed, the participants were asked to complete an online survey. The SMT engines were set to retrain automatically after 200 segments so that the final task would be completed using (lightly) tuned SMT. Participants were requested to wait for three hours between completion of Task 3 and beginning Task 4.

Participant 1 (4990 words)	**Participant 2 (4867 words)**	**Participant 3 (4859 words)**
Condition 1: PE – baseline (100 segments, Test Set 2)	Condition 4: PE with ST capture (50 segments, Test Set 6)	Condition 2: PE with reorder SV (100 segments, Test Set 3)

Condition 4: PE with ST capture (50 segments, Test Set 4)	Condition 2: PE with reorder SV (100 segments, Test Set 1)	Condition 1: PE– baseline (50 segments, Test Set 4)
Condition 3: PE with reorder no SV (50 segments, Test Set 6)	Condition 1: PE – baseline (50 segments, Test Set 5)	Condition 4: PE with ST capture (50 segments, Test Set 5)
Condition 5: PE with incrementally retrained MT (50 segments, Test Set 5)	Condition 6: PE with incrementally retrained MT including condition 2 (50 segments, Test Set 4)	Condition 5: PE with incrementally retrained MT based on condition 2 (50 segments, Test Set 6)

Table 1. Participants' tasks in order

Figure 1 shows the translation editor view on the FALCON portal, with an MT suggestion provided for the current segment. When translating web content, the user can use the Slimview feature to open another browser window, as in Figure 2, to see the website with the current segment highlighted and updated in real time.



Project: shadowcat.skawa.hu_es-ES_myb5oq07_1432053167570.xlf.zip File: shadowcat.skawa.hu_es-ES_myb5oq07_1432053167570.xlf Segments: 1 - 255 Role: Translator

179 Gilles Surprenant, former public works engineer, described it in detail in front of the commission: in ten years, he received from construction companies gifts, invitations to trips, golf tournaments, restaurants, hockey matches and bribes totalling 736,000 dollars, in exchange for sewer contracts of which he inflated the costs. 1 U

Comment:

180 Other highway officials admitted having their palms greased by inflating invoices by 30 to 40%, and by false add-ons. 1 U

Comment:

181 Following these revelations, Mr Tremblay resigned in early November, plunging Montreal into a major crisis. 1 U

Comment:

182 Chantal Rouleau was one of the first women in Montreal to raise the alarm. 1 U

Comment:

183 Mayor of the borough of Rivière-des-Prairies, to the East of the island, she protested in 2010 against the sale of municipal land bought for 5 million dollars and resold for... 1 U

Comment:

184 1.6 million to developers, at the height of the real estate boom. 1 U

Comment:

185 On the investigation which will eventually be implemented, she says she "is following **a thread in order to find out how** the system - infiltrated by ants - works, to put a stop to the gangrene and catch the culprits." | En la investigación que finalmente se aplique, dice ella & ", está siguiendo un hilo para saber cómo el sistema - infiltrados por hormigas - funciona, para poner fin a la gangrena y atrapar a los culpables & ". 1 M

FALCON MT En la investigación que finalmente se aplique, dice ella & ", está siguiendo un hilo para saber cómo el sistema - infiltrados por hormigas - funciona, para poner fin a la gangrena y atrapar a los culpables & ".

Comment:

186 Students, voters considered to be voting more for Democratic candidates, are not allowed in several States to use the photo ID card issued by their institution. | Los estudiantes, los votantes consideran votar más para los candidatos demócrata, no están permitidos en varios miembros a utilizar la tarjeta de identificación con foto emitida por su institución. 1 M

Comment:

187 The process, she says, is "painful but positive." 1 M

Comment:

188 The wound is being cleaned, but Montreal would need its own investigative unit with ongoing monitoring, to avoid the return of these questionable practices. 1 M

Comment:

189 How to clean house. 1 U

Comment:

190 Properly. 1 U

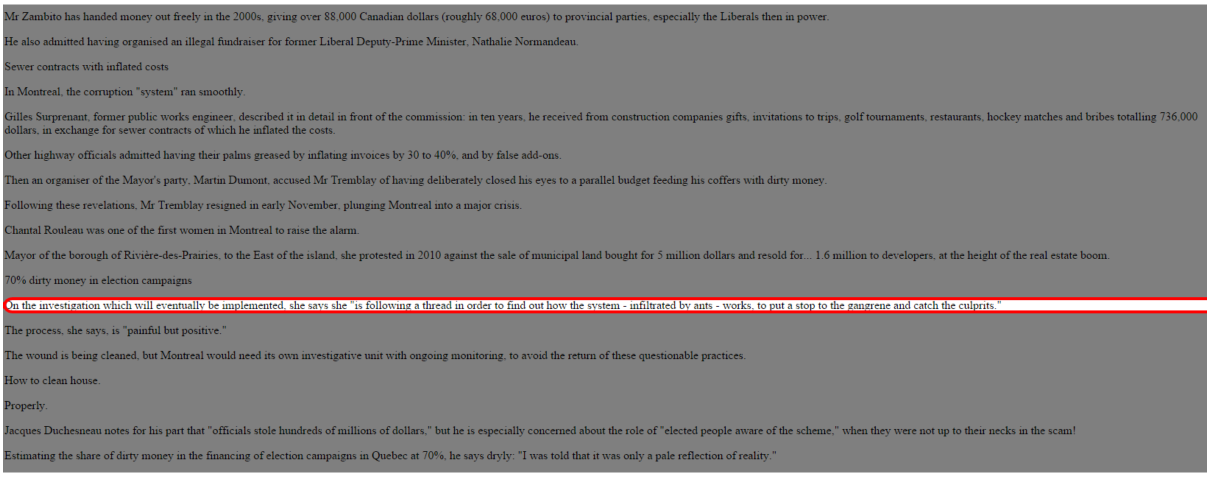
Comment:

191 Jacques Duchesneau notes for his part that "officials stole hundreds of millions of dollars," but he is especially concerned about the role of "elected people aware of the scheme," when they were not up to their necks in the scam! 1 U

Comment:

"We can't go any lower: getting to the bottom of things, with courage, will help to rebuild the house on more solid foundations, with more controls and 1 U

Figure 1. The XTM web interface.



Mr Zambito has handed money out freely in the 2000s, giving over 88,000 Canadian dollars (roughly 68,000 euros) to provincial parties, especially the Liberals then in power.

He also admitted having organised an illegal fundraiser for former Liberal Deputy-Prime Minister, Nathalie Normandeau.

Sewer contracts with inflated costs

In Montreal, the corruption "system" ran smoothly.

Gilles Surprenant, former public works engineer, described it in detail in front of the commission: in ten years, he received from construction companies gifts, invitations to trips, golf tournaments, restaurants, hockey matches and bribes totalling 736,000 dollars, in exchange for sewer contracts of which he inflated the costs.

Other highway officials admitted having their palms greased by inflating invoices by 30 to 40%, and by false add-ons.

Then an organiser of the Mayor's party, Martin Dumont, accused Mr Tremblay of having deliberately closed his eyes to a parallel budget feeding his coffers with dirty money.

Following these revelations, Mr Tremblay resigned in early November, plunging Montreal into a major crisis.

Chantal Rouleau was one of the first women in Montreal to raise the alarm.

Mayor of the borough of Rivière-des-Prairies, to the East of the island, she protested in 2010 against the sale of municipal land bought for 5 million dollars and resold for... 1.6 million to developers, at the height of the real estate boom.

70% dirty money in election campaigns

On the investigation which will eventually be implemented, she says she "is following a thread in order to find out how the system - infiltrated by ants - works, to put a stop to the gangrene and catch the culprits."

The process, she says, is "painful but positive."

The wound is being cleaned, but Montreal would need its own investigative unit with ongoing monitoring, to avoid the return of these questionable practices.

How to clean house.

Properly.

Jacques Duchesneau notes for his part that "officials stole hundreds of millions of dollars," but he is especially concerned about the role of "elected people aware of the scheme," when they were not up to their necks in the scam!

Estimating the share of dirty money in the financing of election campaigns in Quebec at 70%, he says dryly: "I was told that it was only a pale reflection of reality."

Figure 2. The Slimview window.

4. Trial Evaluation Results

4.1. Participant Profiles

Five participants completed the post-editing tasks and the follow-up survey. Of those five, two worked in English to German, two in English to Spanish, and one in English to French. Participants were, on average, in the 41-50 age group, and stated that they had between 5 years and 4 months and 15 years' worth of translation experience. All participants work as freelance translators, with only one stating that they work closely with an agency. Two participants had almost no post-editing experience, one had two years' experience, and two had five years' experience. Only two participants stated that they like to use translation memory (TM) and two feel that the technology is still problematic. Two said that they like to use MT, but they all disagreed with the statements that 'MT helps with my work' and that 'MT is now an advanced technology'. Three participants feel that MT is still problematic. The participants had not worked with XTM before, but commented that they found the interface "user-friendly" and "easier to use than other online translation tools I've used".

4.2. Results

The average baseline speed for participants was 0.228 words per second. Participants spent on average 76.0 seconds on each segment post-editing the baseline MT output. The slowest productivity rate was for the first German participant at 0.136 words per second and the fastest was for the first Spanish participant at 0.303 words per second. As a comparison, in previous research a group of professional post-editors were found to process an average of 0.387 words per second of a text in a familiar technical domain, and a group of novice post-editors processed an average rate of 0.156 words per second (Moorkens & O'Brien, 2015). While SMT output quality tends to be better for French and Spanish, it is difficult to know whether to attribute the varied production speeds to MT quality or to user variance, due to the small number of participants.

Several users found it difficult to download extracted source terms and to upload the file with target terms added. However, once this was completed, this should have made translation more efficient, but the average words-per-second rate was 0.193 using source term capture, and only one participant said that she found the terminological assistance beneficial.

Perhaps predictably, participants were not in favour of post-editing segments out of sequence, with one complaining of the extra difficulty this caused, and another expressing that "it is never a good idea translating texts out of context". They had some difficulty in using the Slimview feature, which may explain why there was little difference in average speed whether they post-edited reordered texts with or without Slimview. The average speed without Slimview was 0.185 words per second, as compared with 0.196 words per second with Slimview.

As had been expected due to the small volume of post-edits used, there was no improvement in productivity after MT retraining. The average speed of post-editing using the output of retrained SMT engines was 0.196 words per second, although one of the participants felt that she perceived an improvement in quality. These figures are tabulated in Table 2.

Participant	**Baseline PE**	**With Source Term Capture**	**Reordered with Slimview**	**Reordered without Slimview**	**PE with retrained SMT**
EN-FR	0.228	0.299		0.248	0.217

EN-DE	0.136	0.144		0.122	0.217
EN-DE2	0.202	0.188	0.157		0.162
EN-ES1	0.271	0.174	0.256		0.174
EN-ES2	0.303	0.159	0.176		0.187
Average WPS	**0.228**	**0.193**	**0.196**	**0.185**	**0.192**

Table 2. Participants' productivity rates for each task (words/second)

4.3. Automatic Evaluation Results

In addition to the participants' productivity tests (Section 4.1), an automatic evaluation was also performed using standard metrics from literature: BLEU (Papineni at al., 2002) and TER (Snover et al., 2006). These metrics compute string-based similarity between the output and a reference translation and give a score between 0 and 1. For BLEU, the higher the score (closer to 1), the better it is. Since TER is an error rate, the lower the score (closer to 0), the better it is.

The evaluation was computed for both post-edited output and machine translated output in terms of BLEU score (Table 3). Since post-editing was performed on the MT output, the expectation is that the scores should be better than the one in brackets (MT output). However there were some discrepancies in the number of sentences evaluated between the two sets mainly due to, (1) The MT system was unable to output certain translations owing to some snags in the web service during the live (time-limited) scenarios; (2) There were snags in recording some of the post-edited outputs via the interface. Both these issues have now been resolved as a result of these studies.

Overall, no significant differences were observed owing to the small size of the output. The TER scores showed similar patterns.

Participant	**Baseline PE**	**With Source Term Capture**	**Reordered with Slimview**	**Reordered without Slimview**	**PE with retrained SMT**
EN-FR	0.51 (0.49)	0.511 (0.49)		0.51 (0.49)	0.512 (0.49)
EN-DE	0.36 (0.357)	0.359 (0.357)		0.359 (0.359)	0.361 (0.360)
EN-DE2	0.361 (0.357)	0.361 (0.357)	0.359 (0.358)		0.36 (0.360)
EN-ES1	0.48 (0.48)	0.48 (0.479)	0.48 (0.479)		0.479 (0.479)
EN-ES2	0.481 (0.48)	0.479 (0.479)	0.479 (0.479)		0.481 (0.479)

Table 3. BLEU scores on Post-Edited output (with MT output in brackets)

4.4. Discussion

Preparation for this evaluation task required the addition of features (such as timing annotation in XLIFF files) and close collaboration between consortium members to remove bugs that only appeared during the pilot and full evaluation. This work was not trivial and caused some delays in carrying out the evaluation, but was ultimately to the benefit of the FALCON platform. At times, problems emerged while participants were part-way through their series' of tasks, interrupting their momentum. This may go some way to explain some of the more curious results as presented in Table 2, such as the decrease in productivity on the part of participant EN-ES2.

However, this work has demonstrated that we have a platform within which users can achieve industrial-speed post-editing with the added integration of terminology tools and proxy web localisation. The successful deployment of automated SMT retraining within the platform places it at the cutting edge in terms of translation technology, and while this did not result in productivity improvements within the task, a perceptible productivity increase would not be expected until several iterations of MT retraining could be completed. Thereafter, the contributions of the post-editor would be statistically more likely to appear in the MT output, and the MT output would in turn become more tuned to the appropriate style and domain.

5. System Security and Intellectual Property Management

5.1. System Security

The FALCON system secures integrated translation and terminology workflows using the existing authentication and authorisation implementations of the XTM, Easyling and Interverbum platforms. XTM additionally acts as the integration hub, offering an OpenID server so that client and LSP personnel can be granted access to the different systems via a single login. The access control features of the individual products are then used to set appropriate permissions for different users, i.e. client project manager, client terminology manager, LSP project manager, LSP translators/posteditors, LSP terminologist. As this is a cloud-based Software-as-a-Service system, third party agency workers can be provided access to the system with the security as for client and LSP staff. The non-user facing systems, such as machine translation, text analytics and L3Data servers are hosted on an XTM server dedicated to the FALCON system and thereby secured using the same server security mechanisms used for other backend servers in an XTM installation. These system security mechanisms are therefore consistent with those already widely deployed for existing XTM, Easyling and TermWeb clients.

5.2. Intellectual Property Management

To achieve the objective of curating 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, e.g. between clients with sensitivities about content leakage or mature TM asset management strategies and LSPs, or between LSP and freelancers or single language LSPs. TMs resulting from work of salaried employees are typically the intellectual property of the employer.

However, in many other 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 vendors and thence to single language vendors 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 task for the parties involved to administer. 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 (as is the focus in FALCON) may have less concern about TM leakage, at least not once the content has been posted. Smaller 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 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[2] and the LetMT! Corpora Repository[3] 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!. This is more challenging to achieve in situations where the publication of parallel text is decentralised by parties without a priori agreements or with variation in the conditions under which the data may be shared and reused. Decentralised publication, has been shown by the linked open data community to support massive scaling in data exchange. However usage rights need to be declared in a way that can be readily indexed and searched, so that parties seeking TM data can quickly determine if specific data is available for the use they need on terms that are agreeable. To this end FALCON has adopted simple, open standard vocabularies, specified in deliverable D2.3, 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 [Bird&Bird]. 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. As discussed above, 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.

The report identifies the following relevant rights that should be addressed in a translation contract. While FALCON, does not aim to provide legal advice on translation contracts it does aim to support the machine-readable declaration of these rights associated with relevant data resources at different points in the lifecycle of data used in translation projects.

[2] https://data-app.taus.net/
[3] https://www.letsmt.eu/Corpora.aspx

Figure 3 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.

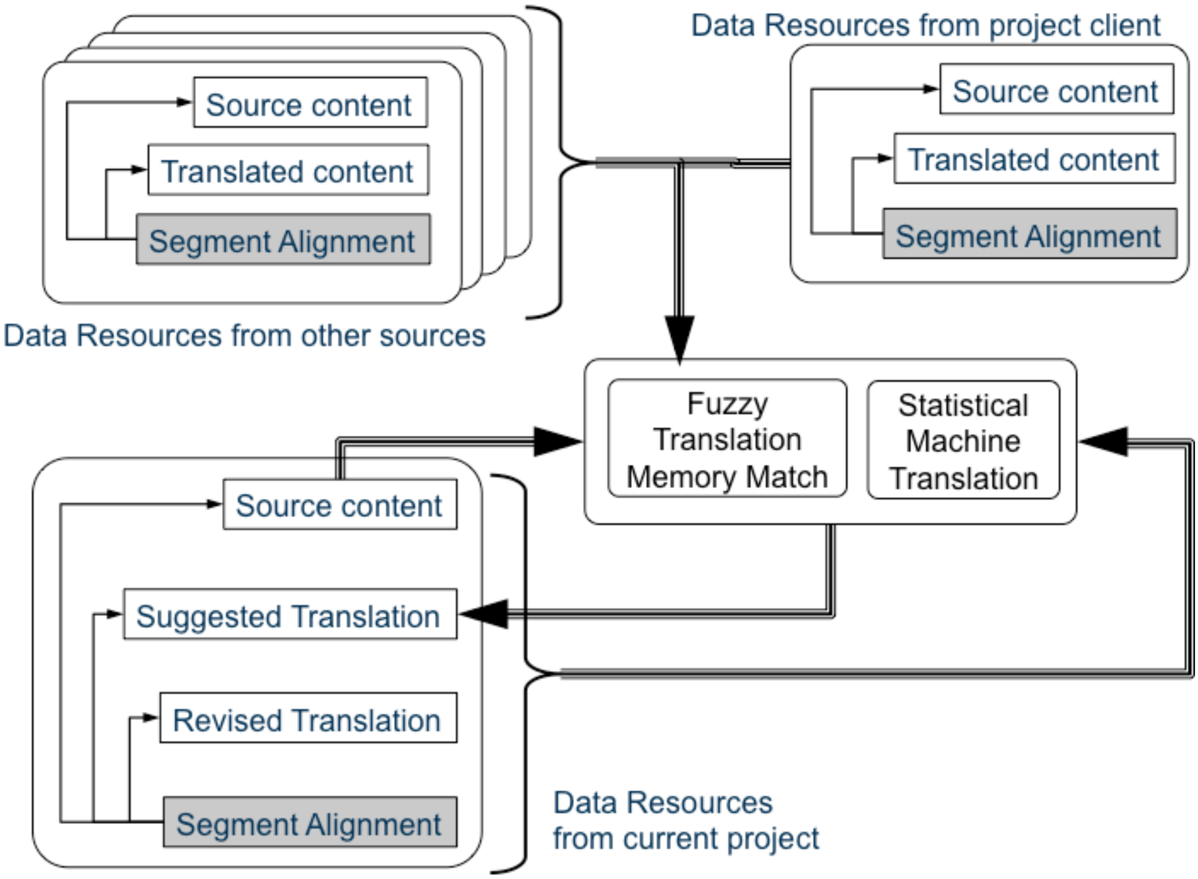


Figure 3. Use of different type of translation data in machine aided translation projects

The relevant intellectual property rights identified in [Bird&Bird] are below summarised against this set of data resources. This summary aims to reflect the main conclusions of [Bird&Bird] in order to assess the sufficiency of the data schema used in addressing the most likely issues. The analysis does not aim to address the many national variations that the report (by its own admission incompletely) identifies, nor the exceptions available for specific content domains, such as scientific works or public sector institute 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:

1. **Data Resources from Client or from other sources**: This consists of:

   1.1. **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. That author also may hold moral rights, whereby their good character may be protected 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.

1.2. **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.

1.3. **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. as TMX, 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 a stronger claim than translators to Sui Generis rights over the alignment.

2. **Data Resources from Current Project**:

2.1. **Source Content**: as for 1.1

2.2. **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.

2.3. **Revised Translations** provided by a human post-editor could be subject to copyright protection as for translated content in general (1.2 above). 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.

2.4. **Segment Alignment** in the project includes the 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:

2.4.1. 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, post-editor and TM database rights holder.

2.4.2. 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. It is unclear however how separate copyright over the source and target documents may impact the use of that content by third parties to generate parallel text using sentence splitting and alignment software. 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.

2.4.3. The alignment between the Source Content and the machine 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.

2.4.4. The segment alignment of the machine Suggested Translation and 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.

2.4.5. With cloud-based translation management systems, such as XTM, it is already common practice to reintegrate Revised Translations into a project TM database so it can be leveraged subsequently in the project. Similarly, FALCON adopts 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.

The L3Data Schema uses open data schema from different standardisation activities related to the W3C Data Activity[4]. 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)[5] which is a W3C Recommendation.
- Dublin Core (dcterm)[6], which is well established meta-data for documents and referenced from DCAT.
- Open Digital Rights Language (odrl), which aims to support machine readable licensing terms – currently a W3C community group document[7]. ODRL in turn, following best practice in open data

[4] http://www.w3.org/2013/data/
[5] http://www.w3.org/TR/vocab-dcat/
[6] http://dublincore.org/documents/dcmi-terms/

vocabulary design, makes use of concepts from other schema, of relevance here is the Creative Commons (cc) vocabulary[8].

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[9]. 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 is 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 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 rule 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 4 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 4 presents the definition for each of the actions and an explanation of how it could be used in the context of translation project data.

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.
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 ‘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.	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 term and condition specify that web analytics and possible A/B testing may be employed in the assessment of translation quality.

[7] http://www.w3.org/ns/odrl/2/ODRL21
[8] http://creativecommons.org/ns
[9] http://www.w3.org/TR/2015/WD-tabular-metadata-20150416/

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[10] 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, 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.
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 applied. This is a common clause in many public TM licenses.
odrl:copy The act of making an exact reproduction of the asset.	This can 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.	Could be use 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 can 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 can 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	This could be used to control extraction of cross-lingual terminology or phrase bi-text from a TM.

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

whether or not the change is significant enough to qualify as a new asset is subjective).	
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.
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 using 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 on a case-by-case basis. May link to a Party with the role function "consentingParty".	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 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).	Can 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.	Can 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 use 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.	Can 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 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.

Table 4. ODRL actions and relevant use in resources relevant to translation

Figure 4 provides the schema used for an L3Data scenario tested in FALCON. 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 accompanies by a CSV-on-the-Web meta-data file[11], which identified the table as a dcat:Distribution and gives the pointer to the relevant ORDL file.

[11] http://www.w3.org/TR/2015/WD-tabular-metadata-20150416/

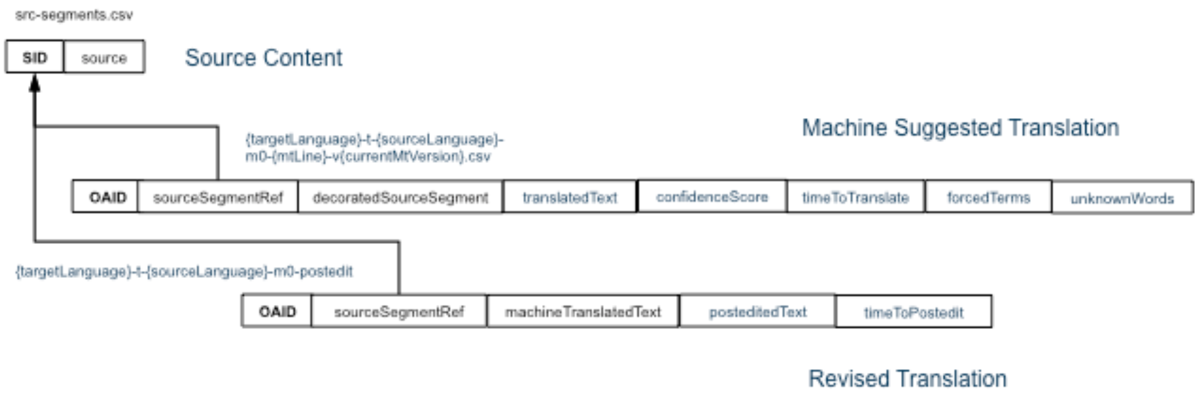


Figure 4. Sample L3Data CSV-on-the-Web Schema

5.3. 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 with 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 4, 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 4) 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 and also circumvention by innovation with technical techniques or argument over the legal interpretation of technical terms. It is therefore recommended that:

  - ODRL be used with domain specific action definitions. This is not prohibited in the current specification, but neither is a mechanism to enable this yet defined.
  - The translation community establish some consensus on action primitive that can be used in ODRL policies that are relevant to its concerns. For example, primitive 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 [Bird&Bird] 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, the FALCON system explored how the management of term lifecycle can be integrated with a project: 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, FALCON explored how third party lexical-conceptual resources such as Babelnet can 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 can be commercially licensed, they depend on a large extent on the aggregation and processing of publically funding or crowd-sourced language resources. While this use is 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. Conversely, FALCON demonstrate how 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 disincentives its capture as reusable datasets. The L3Data scheme provides for separate recording of such term-in-context validation data in a CSV annotation table, and thereby 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 into: the relevant value of term-in-context validation from translation projects in the improvement of text analysis performance and the different uses of text analysis services in the translation industry and more widely in the multilingual web content access industry.

6. References

Translation and Intellectual Property Rights, DGT/2013/TIPRS, July 2014, Bird & Bird LLP

Moorkens, Joss, and Sharon O'Brien. 2015. Post-Editing Evaluations: Trade-offs between Novice and Professional Participants. In Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015), p75-81.

Papineni, Kishore, Slaim Roukos, T.Ward, and W.J. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of Association for Computational Linguistic (ACL 2002), 311–318.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of Association for Machine Translation in the Americas (AMTA 2006), 223–231.