D2.1: REQUIREMENTS SPECIFICATION - REVISED

Mats Granström, Gerd Sjögren, David Lewis

Distribution: Public Report

Federated Active Linguistic data CuratiON (FALCON)

FP7-ICT-2013-SME-DCA

Project no: 610879

Document Information

Deliverable number:	D2.1
Deliverable title:	Requirements Specification
Dissemination level:	PU
Contractual date of delivery:	28th Feb 2015
Actual date of delivery:	5th March 2015
Author(s):	Mats Granström, Gerd Sjögren, David Lewis
Participants:	Interverbum, TCD, XTM, DCU, SKAWA
Internal Reviewer:	DCU
Workpackage:	WP2
Task Responsible:	T2.1
Workpackage Leader:	Interverbum

Revision History

Revision	Date	Author	Organization	Description
1	11/12/2013	Mats Granström, David Lewis	Interverbum, TCD	Initial Draft
2	23/12/2013	Mats Granström	Interverbum	Revised draft
3	6/01/2014	David Lewis	TCD	Revised draft with introductory scenario and MT training use cases.
4	25/1/2015	Mats Granström	Interverbum	Revised in response to reviewer comments, introduced explanatory flow diagrams etc.
5	27/1/2015	Dave Lewis	TCD	Update architecture diagram with TeM-L3D link after tele-conf 26th Jan; Added process flow overview and breakdown figures; Added initial process flow specification text.
6	29/1/2015	Dave Lewis	TCD	Update process flow diagrams and provide sub-process diagrams for G and H after discussion TCD and DCU teams 28/1/2015
7	19/2--25/2 2015	Mats Granström	Interverbum	Added process overview section for mature projects after discussions at Dublin project meeting . Revised use cases and functional details after discussion with project partners.

8	26-27/2/2015	David Lewis	TCD	Added revised process flow based on face to face discussion with Ankit Srivastava, Alfredo Maldonado and John Moran, and remotely with Mats Granström, Patrik Westlund and Andy Zydron
9	28/2/2015	David Lewis	TCD	Implemented corrections from review teleconference with all partners 27/2/2015
10	3/3/2015	Mats Granström Gerd Sjögren, John Judge, John Moran	Interverbum	Removed duplicate information, some text finishing

CONTENTS

1. Executive Summary

The FALCON project combines the power of open data on the web with data-driven language technologies to construct the *Localisation Web*. This consists of a network of terms and translations inter-linked to each other and to source and target documents via URLs. FALCON integrates the resulting web of linked localisation and language data into novel localisation tool chains using existing data models and also query and access control standards. Meta-data from these tool chains add value to these data assets by enabling seamless monitoring of their quality across the value chain and their on-demand leverage in training machine translation and text analytics engines. The interlinked nature of this data will enable publicly available and third party language assets to be dynamically integrated with existing corporate language assets and with the new candidate assets generated by localisation projects. This ensures the continuous collection and curation of targeted, high quality language resources based on the human linguistic judgements exercised within a project, in a process termed *Active Curation*. Within FALCON, we refer to the resulting open data language resources as *Linked Language and Localisation Data (L3Data).*

This document presents a revision of requirements for the FALCON Showcase System, which demonstrates the active curation of language resources as L3Data for (re)training of text analytics and machine translation within a localisation tool chain. This is used to guide the development (WP3) and evaluation (WP4) of the L3Data Federated Platform, its integration with the localisation tools chain and its integration with text analysis and machine translation components. The requirements are structured as: an illustrative business scenario to help communicate the motivation for the requirement, a set of use cases identifying the benefits to specific actors, a detailed process flow capturing the requirements for control flow and data exchange between the system's major components and additional non-functional requirements.

2. Introduction

These are revised requirements for the second cycle of development of the FALCON Showcase System and its constituent components.

These revised requirements are based on prototype development and initial component integration conducted in the first development cycles of the project. They also use knowledge of SME partners in relation to the individual and integrated features that can be offered as extensions to their existing commercial platforms (namely XTM, TermWeb and the Easyling portal). These extensions enable the integration of: language technology (LT) components for statistical machine translation (SMT) and Automatic Term Extraction (ATE); access to open language translation, terminology and lexical-semantic resources; and also end-to-end process management and active curation of LT components using linked data provenance records collected across the tool chain. The requirements are therefore focussed on the tool integration and commercialisation interests of the SME partners in exploiting the integration of the LT components and the L3Data Federated Platform in new solutions for Language Service Providers and Language Service Clients.

An initial high level illustrative usage scenario storyline is presented to give an indication of the different roles that may benefit from use of L3Data and LT integration with tools in existing localisation value chains. This assists in identifying the needs of actors that are then addressed in more detail through a series of use cases.

This is expended into a set of use cases which capture the specific benefits to different actors in the enhanced localisation workflow envisaged by FALCON.

These use cases are mapped onto the system architecture and a process flow model is developed to capture the requirement for data exchange between the components using the integration reference points identified in the architecture. Finally non functional requirements related to the use of interoperability standards, security and exception handling are captures.

3. Illustrative Usage Scenario

The following illustrates a possible high-level usage scenario for the L3Data Federated Platform:

Client Business Setting:
In this usage scenario, we assume that a language service client is running a medical device ecommerce site in the UK. As well as presenting product information, the site offers a user forum where customers may post question and answers on the product range.

Step 1: Localisation Project Requirements
The client initiates a plan to export via its ecommerce site to Poland, France and Sweden and also wishes to find a low-cost, on-going means of translating both the product and user forum content into Polish, French and Swedish.

Step 2: LSP Subcontracting and Project Set-up
The UK institute contacts Easyling to out-source this translation task, since they have handled medical web site translation in the past. The Easyling project manager realises however that the highly specialised nature of the products uses words and phrases not familiar to their usual translators. In addition, the highly perishable nature of the discussion content requires not only fast translation but also commands a low price expectation from the client.

The project manager loads the job into Easyling's subscription to XTM Cloud – L3Data edition. The analytics show a poor TM match against in-house TMs so the project manager realises that targeted use of Language Technology will be required.

Step 3: Terminology Management
To speed the process up, the project manager runs the content through the term extraction system integrated in XTM, connected to TermWeb and a L3Data server, where he has previously stored English medical terminology available from an earlier client. As this termbase had been published by that client as L3Data, the project manager is able to initiate a search for term translated by third-parties in target languages. A terminologist is employed to vet suggested term suggestions and their translations and these are integrated into the project term base in TermWeb.

Step 4: Harvesting Open Corpora (Web Resource Lookup)
The project manager then does an L3Data search on parallel text on medical and health policy domains and finds useful volumes of publicly available bilingual text from the European Community. Downloading this as L3Data via XTM Cloud and combining it with the terminology now available, he generates an initial custom SMT engine, runs the research material through it and distributes the results that need most manual attention to a small team of experienced translators in each target language.

Step 5: Active Curation to retrain SMT engine
Before processing the associated user forum content, the manager retrains the SMT engine using the selected high quality translation, and the text analysis engine with the output of the terminologists stored in TermWeb. Using the analytical tools, the manager notices an appreciable improvement (reduction) in post-editing time, which he confirms using of some automated MT metrics using the new SMT engine. A small sample of QA on the press related materials shows that the shift in terminology usage in the user generated content is not well supported by existing termbase or the results the term extraction from text analysis. However, the translators

working for Easyling also receive small incentives to mark phrases that they come across as being poorly translated frequently, so after a few days of this work, the resulting terms are used to retrain the text analysis engine, alongside another retraining of the SMT engine using selected post-edits.

Step 6: Localisation Process Analytics

Subsequent analysis of post-editing time and SMT metrics show that both the product content and the user posts now being translated on an on-going basis are achieving good enough automated translations that a lighter post-editing regime with less experienced translators can be used for press related material and user posts, enabling reasonable profitability to be maintained on these lower value translations.

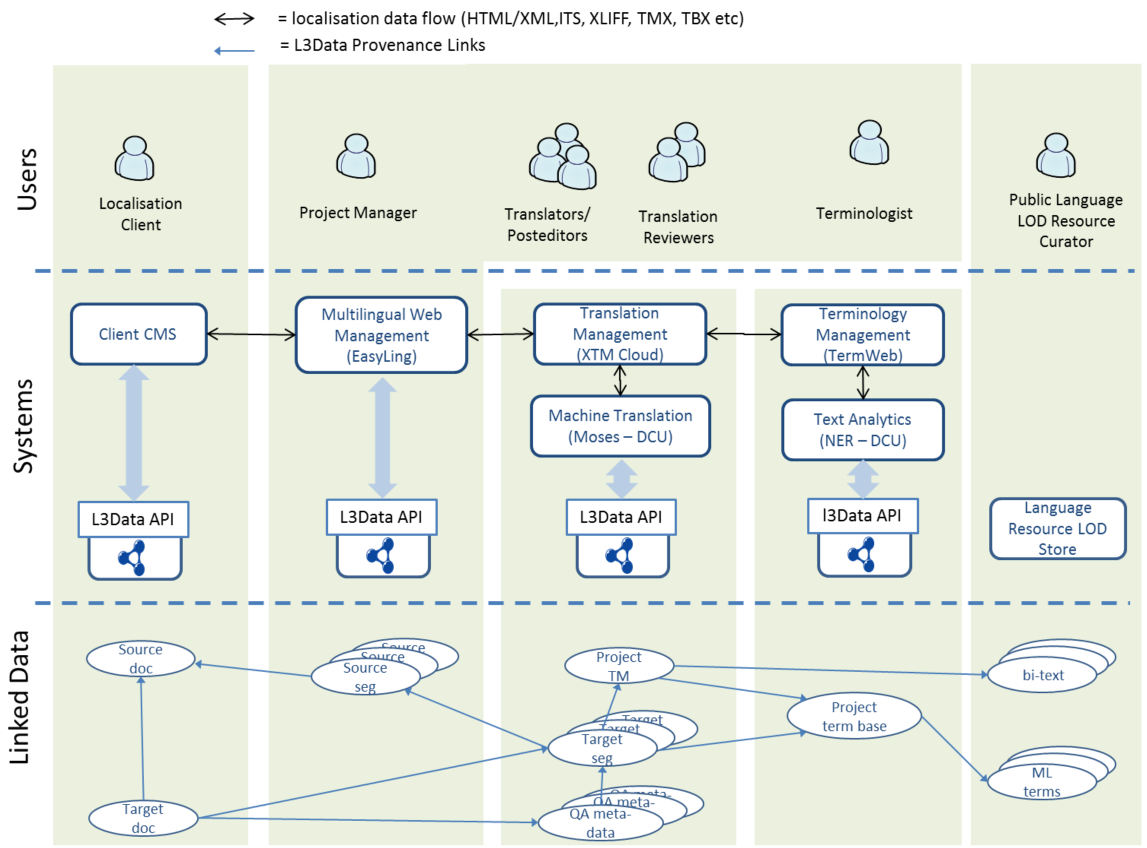
Step 7: Publication of L3Data Generated by Translation Project

The client is pleased with the results as reported back by his European sales and support team. In the project summary report, the client notices the attribution to open linguistic linked data sources in the execution of the project. When asking about this, Easyling explains the benefits of leveraging open L3data and encourages the client to publish the translation memory and termbases from the project, as it may encourage other LSPs and their clients to reciprocate in the future. When the client publishes this data, Easyling annotates it with process quality data. This makes it available to selected LSPs with whom it partners on bigger projects.

Step 8: Third Party Interlink and Reuse of Public L3Data

A Swedish reseller partner of the client is impressed with the term translations resulting from the project, which it can also access via open L3Data. the partner initiates a project with the national health board to further check and QA these translations and then to include them in a new English-Swedish termbase being assembled and published by the local health authorities. The health board is increasingly willing to undertake the cost of such exercises as it sees evidence of the benefits of public term translations published as L3Data in translation projects across its many public facing functions.

Figure 1 below identifies the main system components and the user roles relevant to the L3Data Federated Platform.



= localisation data flow (HTML/XML,ITS, XLIFF, TMX, TBX etc)
= L3Data Provenance Links
Users
Localisation Client
Project Manager
Translators/ Posteditors
Translation Reviewers
Terminologist
Public Language LOD Resource Curator
Systems
Client CMS
Multilingual Web Management (EasyLing)
Translation Management (XTM Cloud)
Terminology Management (TermWeb)
Machine Translation (Moses – DCU)
Text Analytics (NER – DCU)
L3Data API
L3Data API
L3Data API
l3Data API
Language Resource LOD Store
Linked Data
Source doc
Source seg
Project TM
Project term base
bi-text
Target seg
ML terms
Target doc
QA meta-data

Figure 1 Overview of Sample Scenario showing the relationship between users, systems and linked da

4. Use Cases

The following use cases specify the requirements of the system in more detail.

4.1. Use Case A: Automated web site localisation process

Purpose:

Enable automated localisation project and file management for web sites and other content available via the web that needs translation. This will also greatly facilitate the cross-market outreach potential for small and medium-sized enterprises within the EU and reduce the time-to-market for innovative web-based goods and services across the single market and beyond.

Technical details:

1) Easyling engine extracts translatable content from selected web site, packages it and sends it to XTM Cloud.
2) XTM Cloud invokes TermWeb to source integrated terminology, see Use Case B1.
3) XTM Cloud uses customer-specific and federated translation memory resources to provide translation.
4) XTM Cloud calls a statistical machine-translation (SMT) engine, that also has been trained using customer-specific resources as well as federated resources like parallel corpora, to provide machine-translation suggestions, see also Use Case B2.
5) Translator selects, edits, and saves translated segments. Then (selectable, depending on confidentiality requirements of customer) feedback on this editing is automatically sent to the machine-translation engine, thereby retraining it. Here, a semi-automated vetting/approval step is highly recommended. See Use Case B3, D1.

6) Easyling publishes translated web site for customer (on request).
7) Client publishes selected language resources from project, see Use Case E.

4.2. Use Case B1: Seamless integration of prescribed terminology into the translation process

Extends: Use Case A, step 2

Purpose:

Enable automated bilingual lookup of updated preferred company/organisation/project terminology during translation. As only reliable, always updated sources with relevant provenance-tagged data are selected, quick and consistent translations will be provided, even for projects requiring several translators.

Technical details:

1) XTM Cloud to TermWeb: Segment (sentence) to be translated sent to TermWeb that uses a pre-stemming process. Multi-word terms (possibly overlapping) will also be handled, ideally by displaying more than one alternative.
2) TermWeb to XTM Cloud: Segment with term mark-up sent back, including status information etc. (source, definition, part of speech, subject area, term usage status, process status).
3) Translator uses XTM Cloud to view term definition and meta-data and selects and pastes translated term into target text by double-clicking the target-language term.
4) Segment including translation with selected terms stored and used to train the project's connected machine-translation engine.

4.3. Use Case B2: Term extraction, missing terms

Extends Use Case B1

Purpose:

Possibility to (semi-automatically) enhance the termbase using missing terminology

Technical details:

If there is no result for Use Case B1, the project manager initiates a term extraction process from XTM Cloud, resulting in a (source-language) candidate list that the project manager uploads into TermWeb. There, a workflow to supplement terms in the target-language(s) is initiated. During this process the newly created terms will be provided to XTM Cloud for Use Case B1, steps 2 and 3.

4.4. Use Case B3: Suggestion of missing terminology during the translation process

Purpose:

Enable the translator community to contribute necessary terminology

Technical details:

1) XTM Cloud to TermWeb: Translator sends term translation suggestions, corresponding to pre-stemmed source annotations, tagged with term status, preferably annotated with Process status Preliminary.
2) In TermWeb: Translator suggestions taken care of, screened, edited and published via TermWeb workflows. When approved and published they are available in Use Case B1.

4.5. Use Case B4: Term Concordancing

Extends Use Case B1

Purpose:

To enable translators to view the use of the term in a previously translated project.

Technical Detail:

1) Translators in XTM Cloud select a term from the termbase, a term already annotated in a segment or a word or phrase in the segment.
2) A concordance function is called and reference to the prior use of the term in a previous segment is provided.
3) The provenance of the document, e.g. an XML Localisation Interchange File Format (XLIFF) file[1] from which the concordanced segment is taken is also displayed to the translator.

4.6. Use Case B5: Source term extraction and translation based on external resources

Purpose:

To identify possible terms in source text based on external resources that have correct definitions and existing translations that can be provided to improve term translation consistency and accuracy (including disambiguation) in machine translation.

Technical Detail:

1) Possible new terms (not available in TermWeb) are identified by automated term extraction , optionally prioritised by term frequency.
2) External lexical-semantic resources are queried to provide suggested terms, accompanied by definitions, translations, contexts and their provenance.
   Here, the system should be designed to allow for display of multiple (differing) contexts for the same term.
3) Source authors or QA reviewers can confirm or reject suggestions.
4) Additionally, and optionally, terminologists or translators could confirm that the provided translations are appropriate
5) The approved terms and term translations are provided with the content to the SMT engine to ensure consistent translation of terms. This is done through regular calls to TermWeb via its API.
6) Record of the enforced term translation is thus made in TermWeb for access by the SMT engine and also available to translators, who may choose to change this, or to translation QA reviewers. They may choose to annotate this as a translation error. This entails a possibility to store deprecated terms together with the corresponding preferred terms in the termbase.

4.7. Use Case D1: SMT retraining

Purpose:

To retrain an Statistical Machine Translation (SMT) engine during an active project to improve the accuracy of its output to further content within the project.

Technical Details:

1) After a predetermined volume of MT postedits in XTM Cloud have been performed, the project manager selects post-edited text for reuse as training data based on the degree of post-editing performed, the experience of the post-editors and the degree of quality assurance performed on the translations.

[1] http://docs.oasis-open.org/xliff/xliff-core/xliff-core.pdf

2) Once appropriate corpora and term translations have been located, the training corpus is extracted, stripped of inline meta-data as required, normalised, e.g. for language, script and segment length. This occurs prior to being fed into the MT engine training process.
3) The identifier for the new MT engine is annotated with links to the provenance of the training process, the training data selected and the selection criteria used.
4) Relevant customer-approved translation memory segments (paired translations) are transferred to the MT engine automatically when each segment is saved by the translator/proofreader.
5) Regularly update the MT engine with the customer's preferred terminology from TermWeb termbase(s) using API calls.

4.8. Use Case D2: New SMT engine training

Purpose:

- To train an SMT engine based on the selection of corpora published both internally and publically as L3Data.

Technical details:

1) Project managers use L3Data management console to perform a federated query across both internal L3Data sources and external public resources
2) Where public resources return only summary data, the project manager uses this to assess the relevance of the referenced corpora against the project for which the SMT engine is required and its published licensing terms, and on this basis contacts the publisher of the resource to negotiate access to the full corpora or a more detailed summary query.
3) Once appropriate corpora and term translations have been located, the training corpus is extracted. Then prior to being fed into the MT engine for training purposes it is stripped of inline meta-data as required and normalised (e.g. for language, script and segment length,).

4) The identifier for the new MT engine is annotated with links to the provenance of the training process, the training data selected and the selection criteria used.

4.9. Use Case E: Resource Sharing

Purpose:

To allow project managers, either in a language service provider or client organisations, to publish parallel text or terminology corpora under their control for easy discovery and use by other organisations as well as by their own in future projects.

Prepare for public use of linguistic resource (existing or created in project). See Use Cases A and B1 step 4.

Technical details:

1) The system will allow for common methods of accessing and opening up resources for general use. The system will thus provide selection of project-internal or public (federated) resources.
2) Project managers will conduct checks to make sure that no confidential or sensitive information in the source or target content and its associated meta-data is included in public releases. This involves policy driven queries checking for meta-data such as project identification, dates, personnel identifiers, product names etc.
3) The project managers will also have an option to generate statistical summaries of published L3Data corpora which may contain information such as language (pairs), term and word frequency distributions, domain meta-data, MT/PE/QA usage statistics etc.

5. Process Flow Requirements

This section summarises the process flow and data communication requirements for the FALCON Showcase System. This Showcase system is a configuration instance designed to demonstrate the utility and performance of the open interfaces, open data and augmented LT systems in the context of a specific localisation tool chain.

It is intended however that the interfaces and data be sufficiently open to support different process flow configurations and different tool chains..

The Showcase System consists of the following major components:

- Web Site Translation: which accepts submission of web pages from clients and establishes proxy web sites in the requested target languages for the translated version of those web pages. In the FALCON Showcase System this is implemented by an EasyLing Software as a Service (SaaS) instance operated for a Language Service Provider (LSP).
- Translation Management System: which accepts in translatable, extracted and segmented text from the client web site by the Web Site Translation component and returns the translation of this text into the requested target languages. In the FALCON Showcase System this is implemented by an XTM Cloud SaaS translation management and computer assisted translation (CAT) instance operated for an LSP.
- Terminology Management System: which allows the terminology associated with a client and their translation projects, and the translation of that terminology into requested target languages, to be managed. In the FALCON Showcase System this is implemented by a TermWeb SaaS instance operated for the LSP.
- Machine Translation System: which allows Statistical Machine Translation (SMT) engines to be built and retrained for a specific customer and their translation projects. This is offered as a web service managing tailored MT engine instances for the LSP.
- Text Analysis System: which allows Automated Term Extraction (ATE) for a specific customer and their translation projects. This is offered as a web service managing a tailored and automated term extraction instances for the different clients of the LSP.
- Federated L3Data Platform: which allows project specific data for a client to be captured and shared between the different service instances operated by an LSP, for data analyses to support workflow decision-making by the LSP’s translation project managers (PM) and for the publication of language resources generated during a project, under the control of the client.
- Public Data: which are a public source of linguistic data, available for free download, or queryable via a public web service interface.

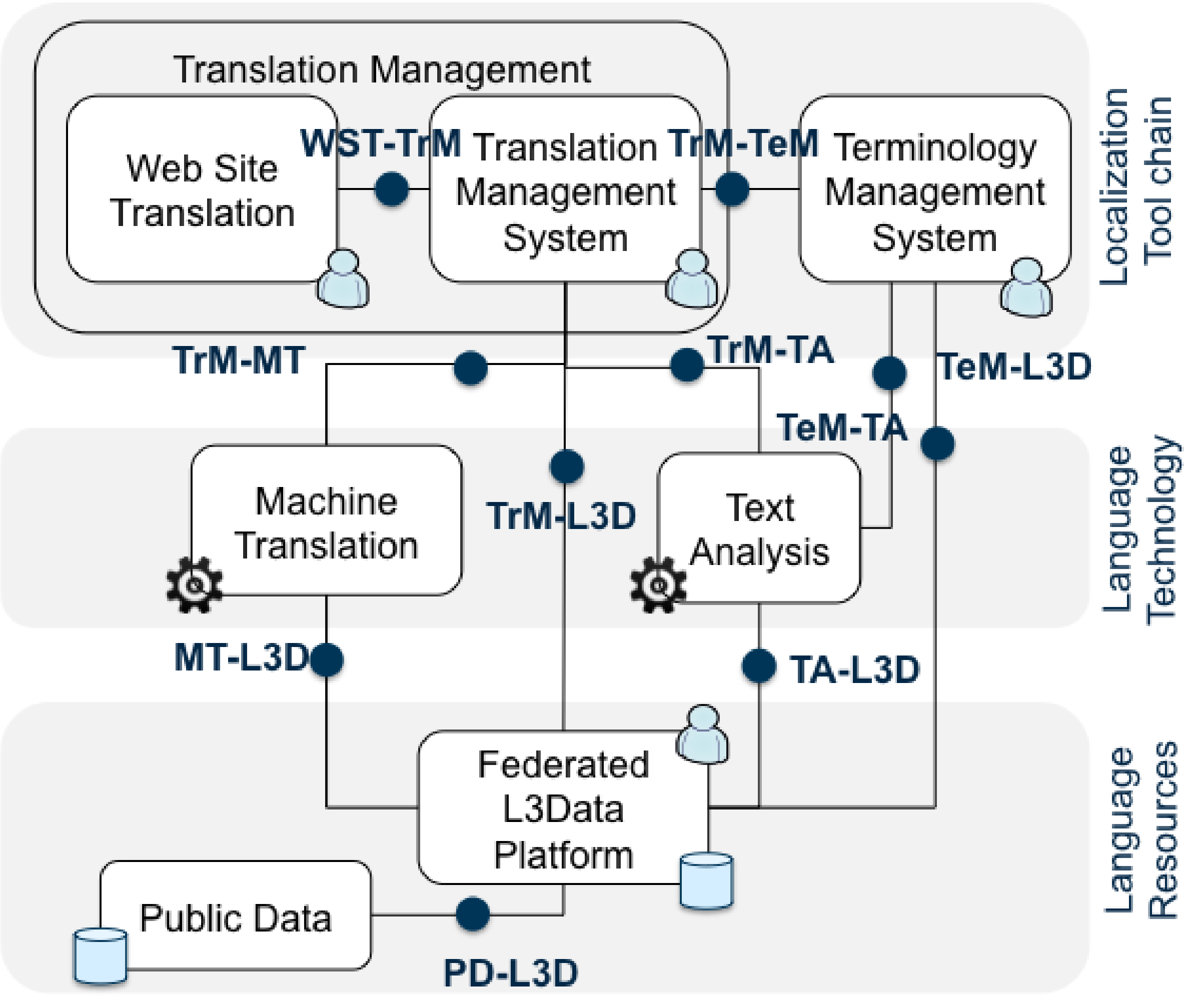


Figure 2: Overview of data flow in the FALCON Showcase System

The process flows below are based on exploration of the capabilities of new or existing interfaces offered by components that may be used at the reference points identified in Figure 2:

- WST-TrM: Easyling to XTM Cloud and vice versa to transfer segmented translatable content and live updates to translations, based on XLIFF. It also supports shared user authentication functions. This functionality is outlined in D3.4 ‘Initial Translation Tool Set Released’ for the first development cycle and will be detailed in D3.8 ‘Revised Translation Tool Set Released’.
- TrM-TeM: XTM Could to Term Web interface where TermWeb decorates words and phrases in a provided segment with meta-data from matches in term base. It also supports shared user authentication functions. This functionality is outlined in D3.5 ‘Initial Terminology Management Tool Set Released’ and will be detailed in documentation for D3.9 ‘Revised Terminology Management Tool Set Released’.
- TrM-MT: XTM Cloud to MT interface where decorated source segments are provided for translation. It also supports requests for MT engine retraining based on provided parallel text. This functionality is outlined in D3.3 Initial SMT and NER components integrated into Platform and D3.4 ‘Initial Translation Tool Set Released’ and will be detailed in the documentation for deliverable D3.7 ‘Revised SMT and NER components integrated into Platform’.
- MT-L3D: Machine Translation to L3Data interface where project source can by submitted for translation to capture initial confidence scores and where the MT component can log output and

confidence scores from later instances of the retrained engine. The retraining of the MT engine can also be triggered via this interface. This is outlined in D3.3 'Initial SMT and NER components integrated into Platform' and will be detailed in the documentation for deliverable D3.7 'Revised SMT and NER components integrated into Platform'.

- TrM-L3D: XTM Cloud to L3Data component. This allows the current state of the translation project in XTM to be accessed as a Translation Interoperability Protocol Package (TIPP)[2] file, containing the source and current target in an XLIFF file. This interface is documented in D3.2 'Initial L3Data Federation Platform Release' and uses an existing XTM API. It also allows an optimised order for segment post-editing to be uploaded to XTM and shared authentication of users, which will be detailed in the documentation for D3.8 'Revised Translation Tool Set Released'. This interface also allows the L3Data component to monitor and advance the project workflow in XTM Cloud through an existing API[3] in order to synchronise it with data analytics and management user interface functions. These features, together with an extension to the TIPP import to access associated resources such as translation memories, term base and user activity logs will be detailed in D3.6 'Revised L3Data Federation Platform Release'.
- TeM-L3D: This allows the L3Data component to access the current start of the project term base including the status of individual terms and their translations. This uses an existing TermWeb API[4], adapted to be synchronised with the term decoration function offer to XTM and with specific rules for managing term status to support validation of automated term suggestions. This is used to inform the optimisation of segment post-editing and target term capture. This will be detailed in D3.6 'Revised L3Data Federation Platform Release' and D3.9 'Revised Terminology Management Tool Set Released'
- PD-L3D: This is the interface to public language resources and is specialised to the access of the BabelNet resources through existing open query API[5] and Babelfy word sense annotation API[6]. Support for importing public Translation Memory eXchange (TMX)[7] and TermBase eXchange (TBX)[8] resources are offered offline. These features will be detailed in D3.6 'Revised L3Data Federation Platform Release'.
- TeM-TA: This interface allows the Text analytics component performing automated term extraction to upload term suggestion to the project term base in TermWeb. It also allows the TA component to check the validation status of these terms as they are processed during the project. This will be detailed in D3.7 'Revised SMT and NER components integrated into Platform'.
- TrM-TA: This interface allows the project workflow to initiate automatic term extraction on the project source. This will be detailed in D3.7 'Revised SMT and NER components integrated into Platform'. it will be mediated by the L3Data component, so this functionality will actually be supported via the TrM-L3D interface.

As indicated above the detailed specification of individual interfaces is distributed over deliverable associated with the delivery of the corresponding software implementations, however this documentation will be consolidated into a single reference document to be released with D2.3 'Final L3Data Schema and Architecture'.

[2] http://interoperability-now.googlecode.com/files/The_TMS_Interoperability_Protocol_Package-1.5.pdf
[3] http://www.xtm-intl.com/manuals/XTMConnectSDK.pdf
[4] http://www.termweb.org/docs/api/
[5] http://babelnet.org/guide
[6] http://babelfy.org/download.jsp
[7] http://www.gala-global.org/oscarStandards/tmx/tmx14b.html
[8] http://www.tbxinfo.net/

5.1. Overall Process Flow - new customer and project

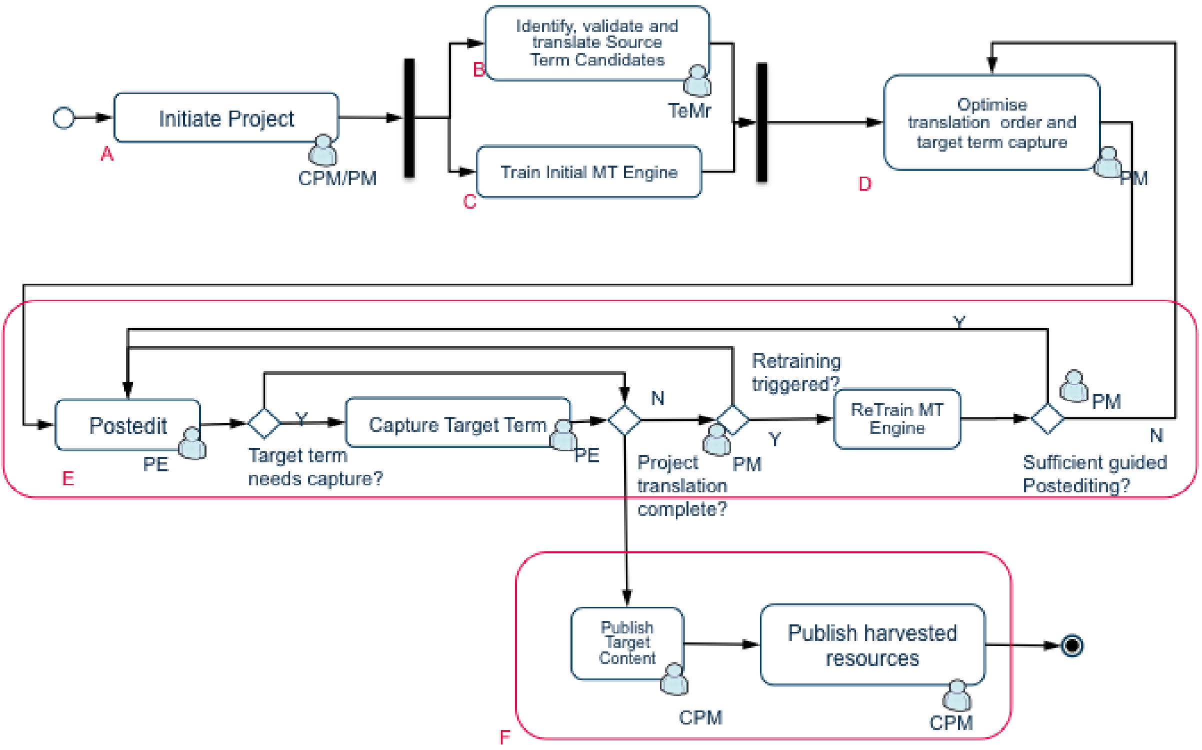

Initiate Project
A
CPM/PM
Identify, validate and translate Source Term Candidates
B
TeMr
Train Initial MT Engine
C
Optimise translation order and target term capture
D
PM
Postedit
E
PE
Y
Target term needs capture?
Capture Target Term
PE
N
Project translation complete?
Retraining triggered?
PM
Y
ReTrain MT Engine
Y
PM
N
Sufficient guided Postediting?
Publish Target Content
CPM
Publish harvested resources
CPM
F

Figure 3. Process flow for new projects

The process flow addresses a use case where a new web site has been established and where the web site owner, i.e. the client, does not have pre-established terminology or translation memories that can be leveraged in the project. They request that an LSP translate the web pages as quickly, accurately, consistently and cheaply as possible. This is done though the following process steps:

A) The client project manager sets up an account in Easyling and initiates a project, and corresponding client and project level configurations are established in the federation of systems that makes up the FALCON Showcase System instance as operated by the LSP

B) Automated terminology extraction is used to identify possible terms. TermWeb is used to validate these terms and provide translation where possible. This makes use of Public Data as well as existing public term bases that are available to the client's project via TermWeb. Validations are used to improve the automated term extraction component so that it works more accurately for future project by the same customer or in similar domains.

C) In parallel, an initial SMT instance is prepared for the project using public parallel text.

D) The LSP PM conducts an analysis to optimise the order for post-editing and accompanying capture of term translations. For the initial portion of this optimised segment sequencing, more experienced post-editors are used in a guided manner.

E) Post-editing commences guided by optimised segment sequence. The disruption to the understanding of context by post-editors is mitigated by a synchronous view of the current translated segment in context via the WST. Post-editors are also encouraged to capture term translations and these are used to provide consistent term translation by the SMT component for subsequent segments. The LSP PM periodically authorises the retraining of the SMT instance, based on the post-edits provided. The LSP PM monitors the guided post-editing process based on post-editing performance measures (time to post-edit and regressive automated MT assessment scores). The PM decides if the improvement in SMT output is sufficient to justify switching to a straight sequential translation order, with post-editing possibly undertaken then by less experienced translators.

F) On completion of the translation, the client PM is able to publish the translations on the target market language web site proxy and also make decisions on what terminological and parallel text generated by the project may be published and under what licenses or constraints.

The requirements analysis arising from the process flow is broken down according the sub-processes areas A to F in figure 3 and these are detailed in the sections below.

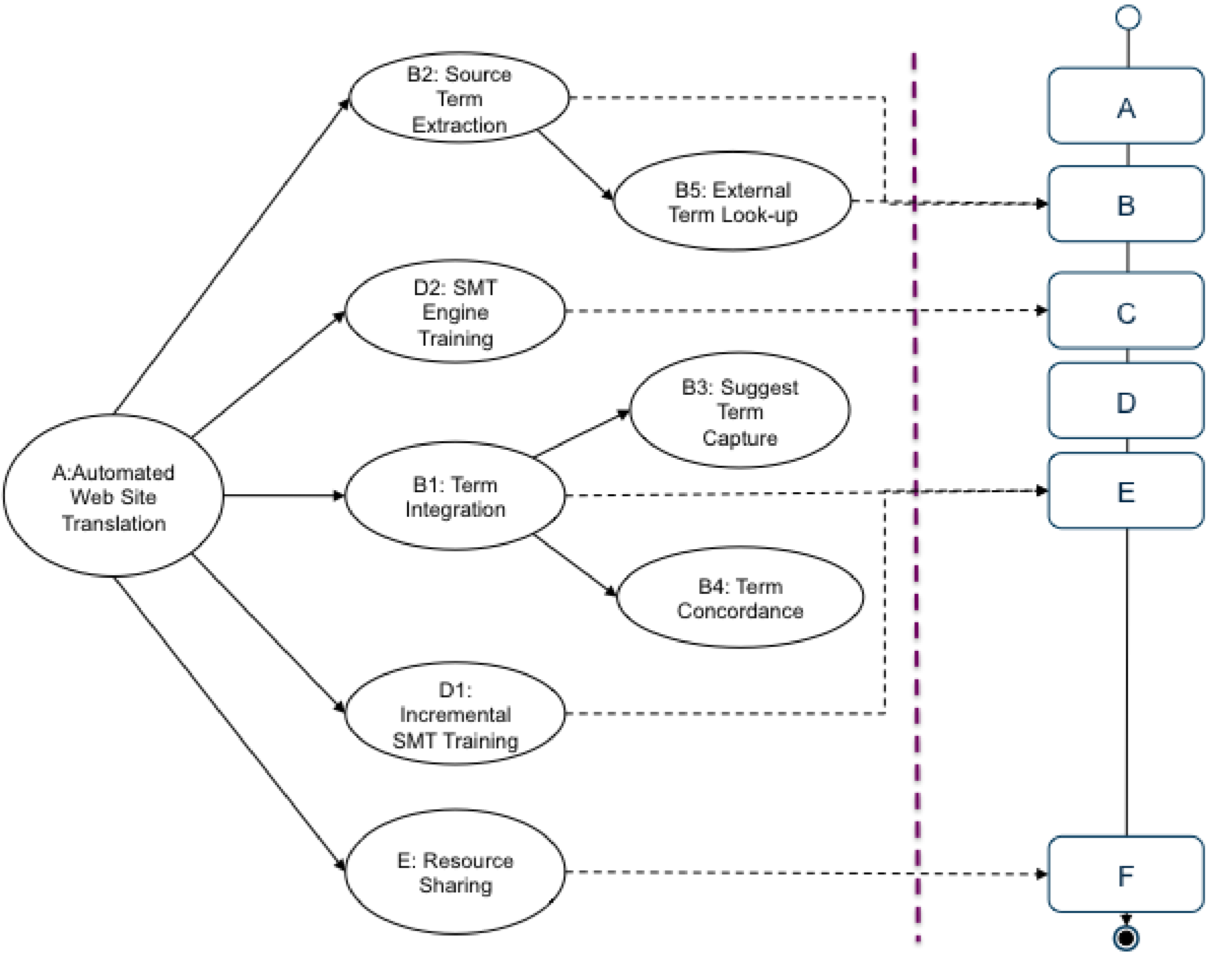


Figure 4: Mapping of use cases to sub-processes

Figure 4 provides a mapping between the use cases in section 4 and the sub-processes defined in this section. Note that the process flow represents the subset of the use cases that most directly demonstrate the active curation of L3Data and its use in retraining Lt components. While other features of the translation and term management tool chain may also potentially interact with the L3Data , e.g. term concordancing and TM lookup, these are not expanded upon in the sub-processes, and they are already available in a form within the tool chain, and do not contribute directly to retraining.

5.2. Sub-process A: Initiate Project

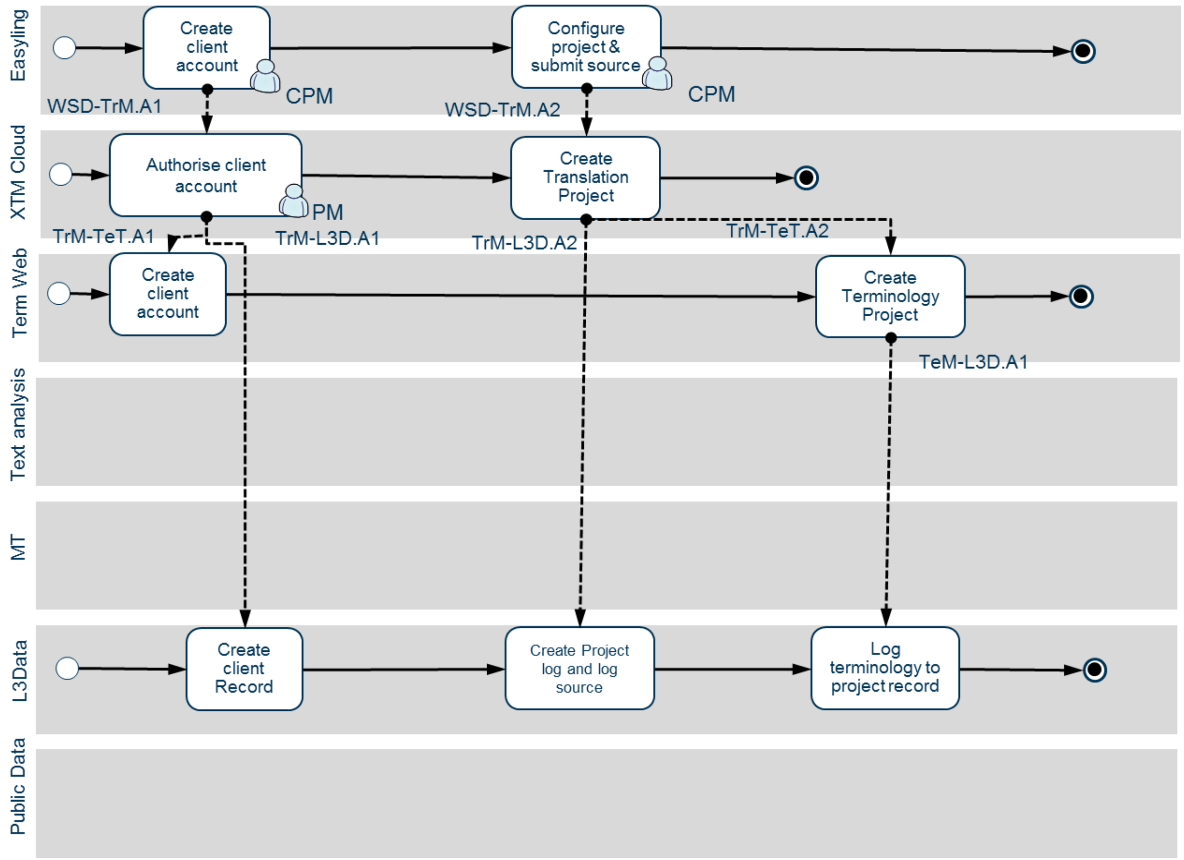


Figure 4: Sub-process A: Initiate Project

ACTIVITY Create client account

- LSP Client PM (CPM) requests an account for FALCON system with themselves as Client Business Contact.
- WSD-TrM.A1: LSP PM (hereafter PM) sets up an account in XTM Cloud for the client, which in turn gives access to client account in Easyling, TermWeb and L3Data component through preconfigured federated authentication for PM role:.
  - INPUT:
    - Client company identifier
    - Client PM Role credentials (user ID and password)
  - BEHAVIOUR:
    - Actor: LSP Project Manager
    - ACTIVITY: Authorise Client Account in XTM and other modules
    - TrM.TeM.A1: PM requests TermWeb client account (LSP has dedicated TermWeb instance).
      - INPUT:
        - Client company identifier
        - Client PM identifier and reference for authentication
        - LSP PM identified as LSP PM for this client
      - BEHAVIOUR: TermWeb client account established and CPM and LSP PM

added as roles

- RETURN:
  - Session ID for management of client.
- TrM.L3D.A1: XTM Cloud project manager requests client account in L3Data Platform
  - INPUT:
    - LSP Name
    - LSP PM Role credentials
    - Client company identifier
    - Client PM identifier and reference for authentication
    - LSP PM identified as LSP PM for this client
  - BEHAVIOUR:
    - Create Actor record for client, for client PM and for LSP PM associated with this project
  - RETURN:
    - Base URL for REST interface for interaction with L3Data for this client
- RETURN:
  - Confirmation of client account set up

ACTIVITY: Specify Translation Project & Submit Source

- Client PM sets up project parameters and submits source. The base scenario presented here assumes vanilla client with no prior Translation Memory or Term Base, however if these exist they would be uploaded here, or included from previous projects conducted by this LSP.
- WSD-TrM.A2: Easyling requests project set up in XTM Cloud
  - INPUT:
    - Client credentials
    - Project Name
    - Target language(s) (included in XLIFF)
    - Delivery Date
    - XLIFF File containing source
  - BEHAVIOUR: Create translation project in XTM
    - ACTIVITY: Set up project details in XTM Cloud with information provided, selecting a workflow that includes automated term extraction, MT with term forced decoding and iterative retraining. Configure project in TermWeb and L3Data components.
    - TrM.TeM.A2: XTM Cloud project manager requests project set-up in TermWeb
      - INPUT:
        - Client session ID
        - PM credentials
        - Project name
        - Optionally, select relevant TermWeb dictionary or dictionaries will be stated in XTM when setting up the project.
      - BEHAVIOUR:
        - Set up dictionary for project including existing relevant general purpose or client specific dictionaries.
      - RETURN:
        - Dictionary ID for client project.
    - TrM.L3D.A2: XTM Cloud project manager requests project set-up in L3Data Platform
      - INPUT:

  - LSP PM credentials
  - Project Name
  - TIPP containing XLIFF with source and project parameters in Linport STS format specifying:
    - Production task plan, e.g. client-term-capture; auto-term-extract; term-validation; term-translate; mt-generation; mt-translate; mt-postedit; postedit-term-capture; publish content; publish data
    - Source content
    - Target Languages
    - TermWeb Dictionary ID for client project
- BEHAVIOUR:
  - Create a collection to record logs associated with this project, recording project parameters and plan taken from STS and XLIFF files
  - Create a log entry for reception of source content pointing to table containing source content indexed by XLIFF segment ID
- RETURN:
  - REST URL for access all LSP L3data, associated with project, e.g. www.l3d.com/id-thislsp-001/proj-001

- RETURN:
  - ProjectID

5.3. Sub-process B: Identify and Validate Source Term Candidates

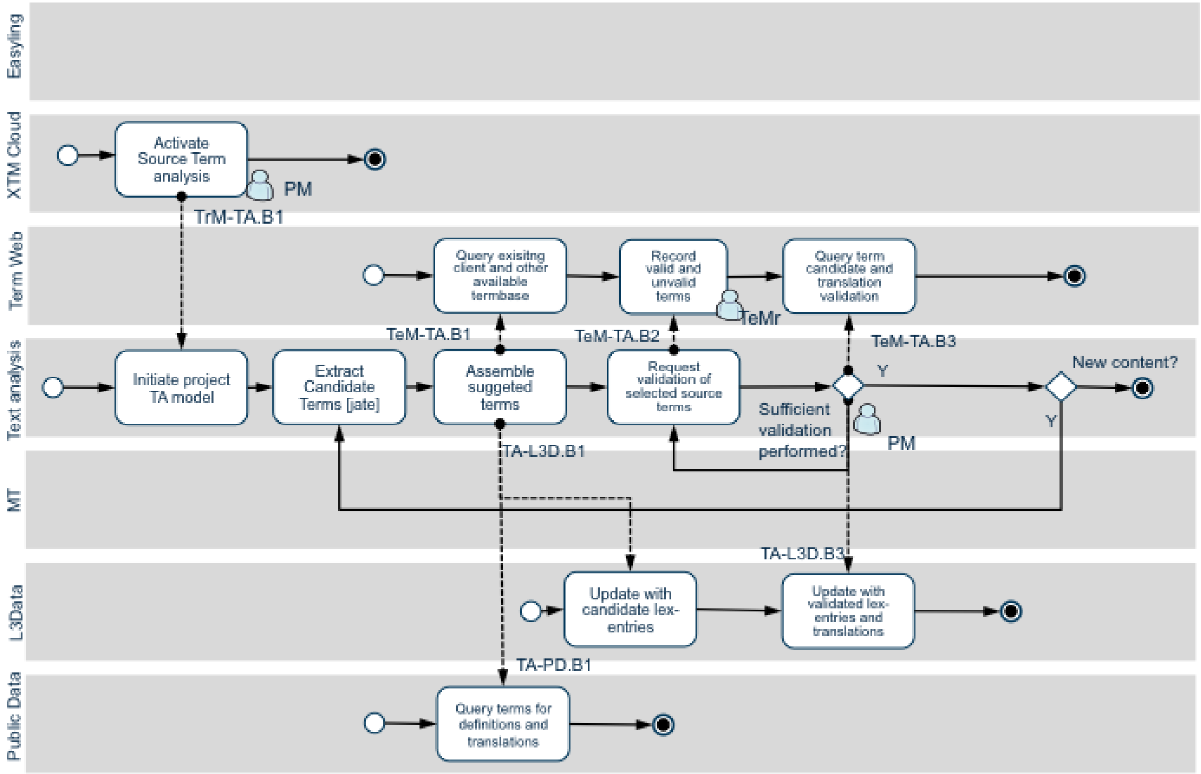

Easyling
XTM Cloud
Term Web
Text analysis
MT
L3Data
Public Data
Activate Source Term analysis
PM
TrM-TA.B1
Query exisitng client and other available termbase
Record valid and unvalid terms
TeMr
Query term candidate and translation validation
TeM-TA.B1
TeM-TA.B2
TeM-TA.B3
Initiate project TA model
Extract Candidate Terms [jate]
Assemble suggeted terms
Request validation of selected source terms
Sufficient validation performed?
PM
Y
New content?
Y
TA-L3D.B1
TA-L3D.B3
Update with candidate lex-entries
Update with validated lex-entries and translations
TA-PD.B1
Query terms for definitions and translations

Figure 6: Sub-process B: Identify and Validate Source Term Candidates

ACTIVITY: LSP Project Manager Initiates Source Term Analysis:

TrM-TA.B1: XTM Cloud Requests automated source term extraction from Text Analysis Service

- INPUT:
  - LSP project manager credentials
  - L3Data REST URL for project
- BEHAVIOUR:
  - If client or project L3Data term log does not exist create one
  - Run automated term extraction on project source to get list of candidate terms based on criteria set by LSP project manager, e.g. match to client-project termbase, cut off for term frequency
  - ACTIVITY: Assemble list of suggested terms from automatically extracted terms ->
  - TeM-TA.B1:
    - INPUT:
      - Suggested term list
    - BEHAVIOUR: Query for matches of suggested terms extracted from source with client project terms base
    - RETURN
      - Term specifications for matches, including translations
  - TA-PD.B1

    - INPUT:
      - Suggested term list
    - BEHAVIOUR: Query for matches for automatically extracted terms from public lexical-conceptual resource, e.g. BabelNet
    - RETURN
      - Term definitions and translations
  - ACTIVITY: Validate suggested source terms:
  - TeM-TA.B2:
    - INPUT
      - Consolidated list of suggested terms
    - BEHAVIOUR: Human validation of term definitions and term translations as well as association of terms into concepts, including pre-existing concepts in client project term base
    - RETURN
      - TermWeb Client Project dictionary ID
  - ACTIVITY: Analyse validation of suggested source terms and their translation to date, including which are validated, which were rejected and which had not been reviewed to date:
  - TeM-TA.B3:
    - INPUT
      - TermWeb Client Project dictionary ID
    - BEHAVIOUR: Query for validation, non-validation and not yet reviewed for suggested term in Termbase
    - RETURN:
      - Term, concept and translation status from TermWeb client project termbase.
  - TA-L3D.B3:
    - INPUT
      - LSP PM Credentials
      - L3Data client project URL
    - BEHAVIOUR: Log updates to term validation
    - RETURN:
      - REST URL of term validation log
- RETURN: If LSP PM satisfied with degree to which suggested terms have been validated then
  - REST URL to access validated client-project L3Data terminology suggestion log.

If new content is subsequently submitted by the client , the validated and non-validate term suggestions are used to improve the accuracy of the automated term extraction component.

5.4. Sub-process C: Train Initial MT Engine

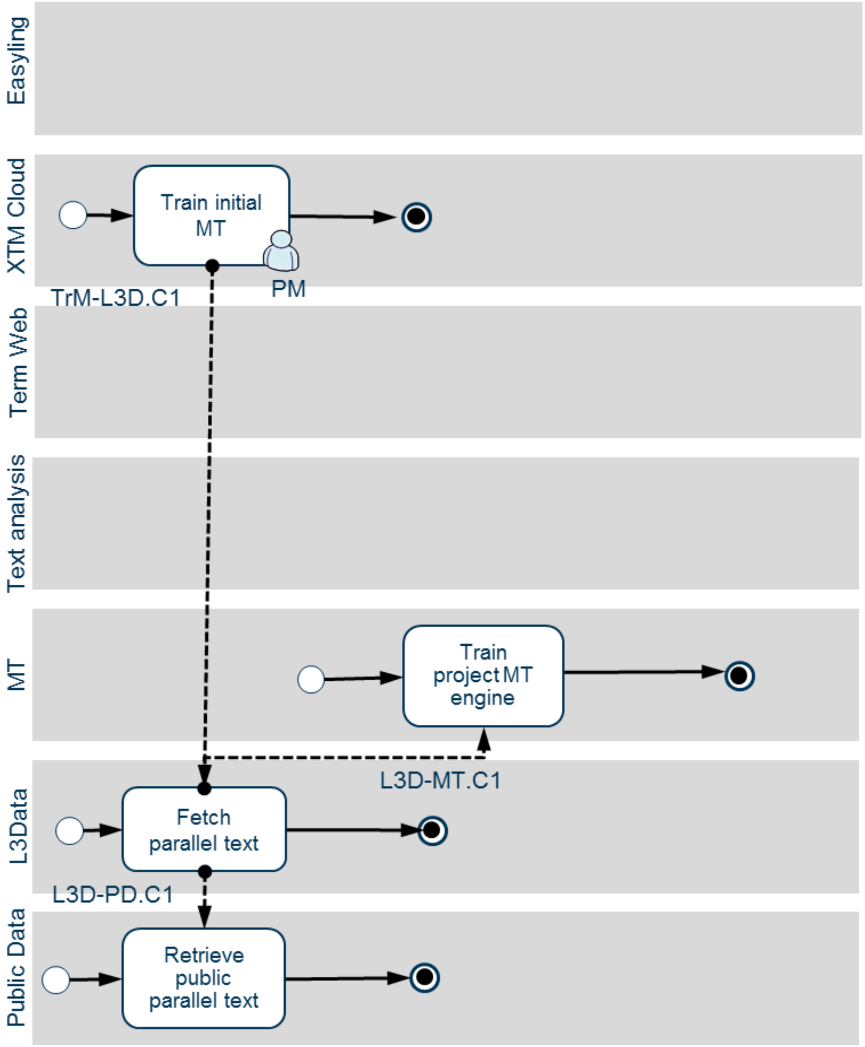


Figure 7: Sub-process C: Train Initial MT Engine

ACTIVITY: Train Initial MT Engine: Management Workflow step triggered to initiate training of initial MT engine.

TrM-L3D.C1: Management workflow step to initiate

- INPUT:
  - PM credentials
  - Client ID and Project name
  - Language pairs for required project MT engines
- BEHAVIOUR:
  - ACTIVITY: Retrieve Public Parallel Text
  - L3D-PD.C1
    - INPUT
      - Required Language pairs
      - Optionally: domain information

- - - BEHAVIOUR: LSP PM is offered a summary of public training data available. If client-provided training data, or the LSP has training data it can reuse in this project, these will be included also. The LSP PM can approve the selection of training data.
      - RETURNS
        - Reference to approved parallel text.
  - ACTIVITY: Train project MT engine
  - L3D-MT.C1
      - INPUT:
        - LSP PM credentials
        - Client ID and project name
        - Required language pairs
        - Reference to parallel text for each language pair
      - BEHAVIOUR: for each language pair, generate a separate MT engine using the provided parallel text.
      - RETURNS
        - Web Service URL for individual MT engines
        - REST URL for training log record
- RETURNS:
  - Web Service URL for individual MT engines

5.5. Sub-process D: Optimise Post-edit Order and Target Term Capture

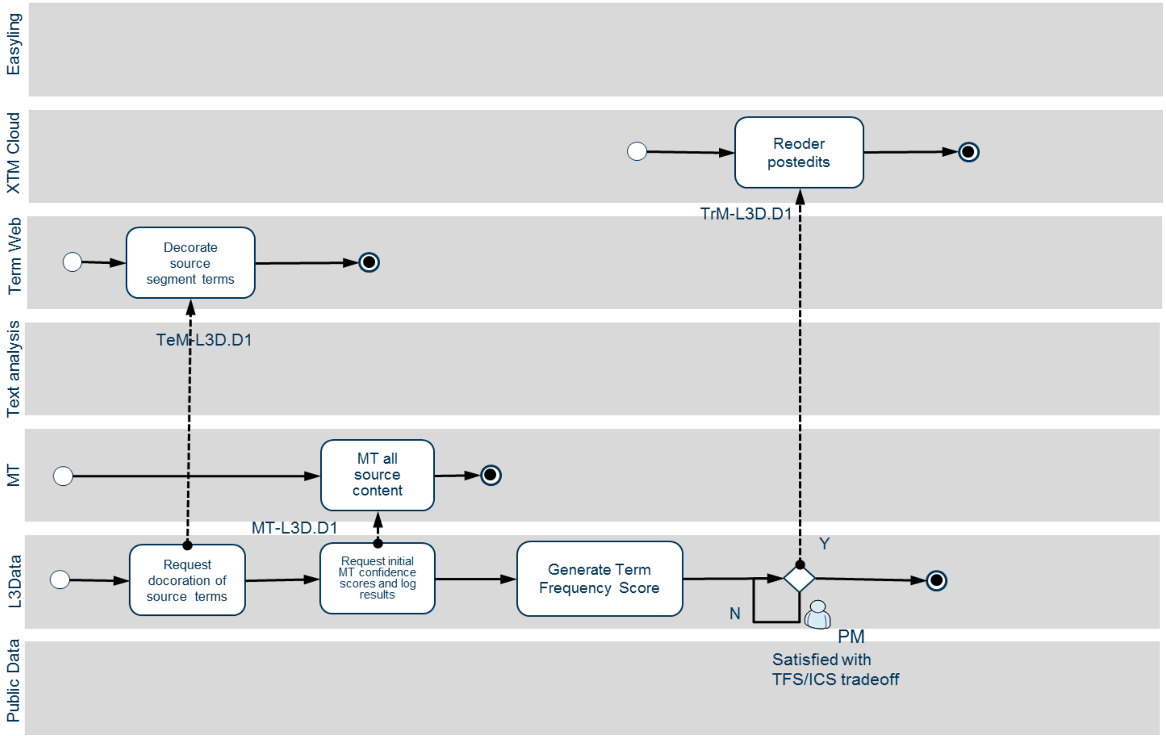

Easyling
XTM Cloud
Term Web
Text analysis
MT
L3Data
Public Data
Reoder postedits
TrM-L3D.D1
Decorate source segment terms
TeM-L3D.D1
MT all source content
MT-L3D.D1
Request docoration of source terms
Request initial MT confidence scores and log results
Generate Term Frequency Score
Y
N
PM
Satisfied with TFS/ICS tradeoff

Figure 8: Sub-process D: Optimise Post-edit Order and Target Term Capture

ACTIVITY: LSP PM optimises post-editing order and target term capture

- TeM.L3D.D1: Request decoration of whole project source with current validated term identification where valid term translation exist in each target language
  - INPUT:
    - LSP PM credentials
    - Client ID and project dictionary ID
    - Source text
    - Target language pairs
  - BEHAVIOUR: Decorates each segment where each term matches a validated term and add term translation for each target language
  - RETURNS:
    - Term decorated segments
- MT-L3D.D1:
  - INPUT:
    - LSP PM Credentials
    - Client ID and project dictionary ID
    - Source text decorated with source terms
    - Target language pairs
  - BEHAVIOUR: Translate each source segment, forcing corresponding output for decorated segments with provided translation
  - RETURNS:
    - Target text and accompanying translation confidence score

- ACTIVITY: Generate Term Frequency Score, indicating for each segment a measure of the frequency of terms in that segment across the job, for terms where no validated translation exists
- ACTIVITY: LSP PM set post-editing order by trading off between prioritising post-editing of segments with worst machine translation output (based on confidence score) and with the most important terms that still require valid translations. This trade-off should reflect business level factors including the post-editing expertise and target term capture expertise of available post-editors; the relative costs of post-editors with different levels of expertise; the prevalence of poor MT output and un-validated term translations across the project and for different languages.
- TrM-L3D.D1: Configure
    - INPUT:
        - LSP PM credentials
        - Client ID and project name
        - Reorder segment IDs for each target language
    - BEHAVIOUR: Sets order in which segments are presented for post-editing in each target language and may select to assign early order to more experienced translators

5.6. Sub-process E: Guided Post-editing and Target Term Capture

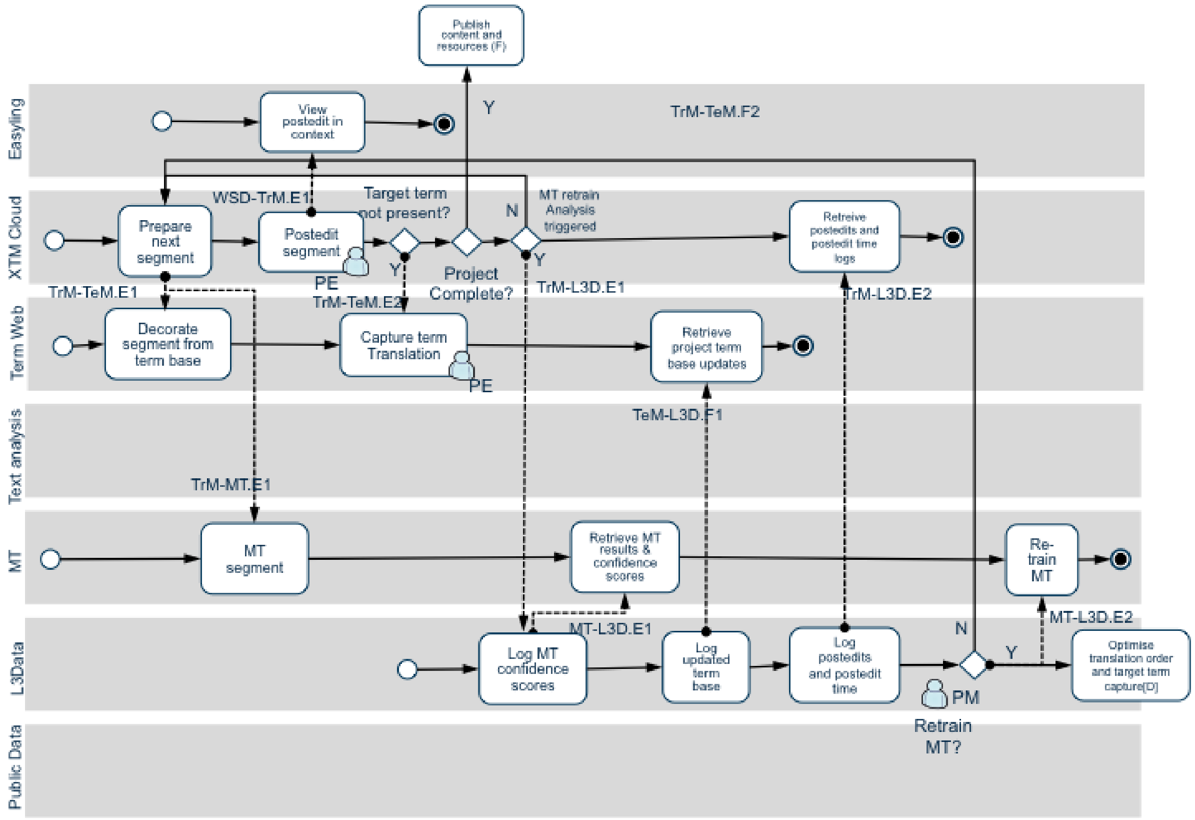


Figure 9: Sub-process E: Guided Post-editing and Target Term Capture

ACTIVITY: Process each project segment in the order specified by the post-editing and target term capture optimisation analysis. For each segment:

- TrM-TeM.E1: Decorate Segment source with terms
    - INPUT
        - Post-editor (PE) credentials
        - Client ID and project name
        - Source Segment
        - Target Language
    - BEHAVIOUR: Decorate segment by annotating words or phrases that, once stemmed, match an entry in the project term base. The decoration includes the translation of the term if present.
    - RETURNS
        - Decorated segment
- TrM-MT.E1: Machine translate segment
    - INPUT
        - LSP credentials
        - Client ID and project name
        - Segment(s) to translate with term decoration
        - Target language
    - BEHAVIOUR: Translate each source segment, forcing corresponding output for decorated

segments with provided translation. Logs confidence score internally.

  - RETURNS
    - Translated segment(s)

- ACTIVITY: Assigned post-editors post-edit segment in XTM Cloud CAT function
- WSD-TrM.E1: Post-edit segment in context
  - INPUT
    - Segment ID
  - BEHAVIOUR: Post-editor may view currently post-edited segment in Easyling HTML page viewer in the target language, which is populated with current MT output or results of post-editing for each segment on the page. The segment currently being post-edited is highlighted and can be edited in the in context view.
  - RETURNS
    - Completed segment post-edit
- ACTIVITY: If a term decorated in the source is not decorated with a translation, then the post-editor may invoke TermWeb and enter the translation they have provided when post-editing the segment.
- TrM-TeM.E2: TermWeb is opened within a segment post-editing session
  - INPUT:
    - PE credentials
    - Client Id and dictionary ID
  - BEHAVIOUR: Post-editor adds the translation for the decorated term to the term entry.
  - RETURNS: No direct response, but the added term translation is used therein to decorate terms.
- ACTIVITY: If this is the last remaining segment of the project, workflow advances to Publish Content and Data
- ACTIVITY: After a specified number of segment post-edits are completed in the project for a specific target language, an MT retraining analysis is triggered
- TrM-L3D.E1: Signals collection of L3data needed to conduct MT retraining analysis
  - INPUT:
    - LSP PM credentials
    - Client ID and project name
    - MT instance ID
  - BEHAVIOUR: Initiate collection of data for MT retraining analysis
  - MT-L3D.E1: Log MT Confidence Scores
    - INPUT:
      - LSP PM credentials
      - Client ID and project name
      - MT instance ID
    - BEHAVIOUR: Retrieve log of segments translated by MT engine instance including term translations used in decoding and segment translation confidence score
    - RETURN
      - Reference to MT instance translation logs
  - TeM-L3D.E1: Log updated term base
    - INPUT:
      - LSP PM Credentials
      - Client ID and Dictionary ID
    - BEHAVIOUR: Retrieve updates to term web from post-editing process, specifically newly added term translations and term validations
    - RETURN:

        - Reference to Termbase updates
- TrM-L3D.E3: Log post-edits and per-segment post-edit time
    - INPUT
        - LSP PM Credential
        - Client ID and Project Name
    - BEHAVIOUR: Retrieve record of post-edits of machine translation and time log of per-segment post-editing
    - RETURN
        - Project TIPP file containing XLIFF with MT suggested translation and user activity data capturing time to post-edit
- ACTIVITY: LSP PM assesses whether to authorise a retraining of the MT engine, using gathered post-edits to produce a further MT engine iteration. This is based on analysos of: changes observed in the post-editing time using previous iterations; the relative cost of post-editors with different levels of experience; the completion of translations of validated terms and the proportion of the project completed. If the LSP PM opts not to retrain the MT at this point, the project continues to the next segment for post-editing. If they do opt to retrain then the following applies:
- MT-L3D.E2: Request retraining of MT engine
    - INPUT:
        - LSP PM credentials
        - Client ID and project name
        - Reference to training data
    - BEHAVIOUR: MT is retrained by integrating the new training data into the engines model
    - RETURN
        - ID of new version of MT engine
- ACTIVITY: Reanalyse the optimisation of post-editing and target term capture before proceeding to post-edit the remainder of the project

5.7. Sub-process F: Publish Content and Data

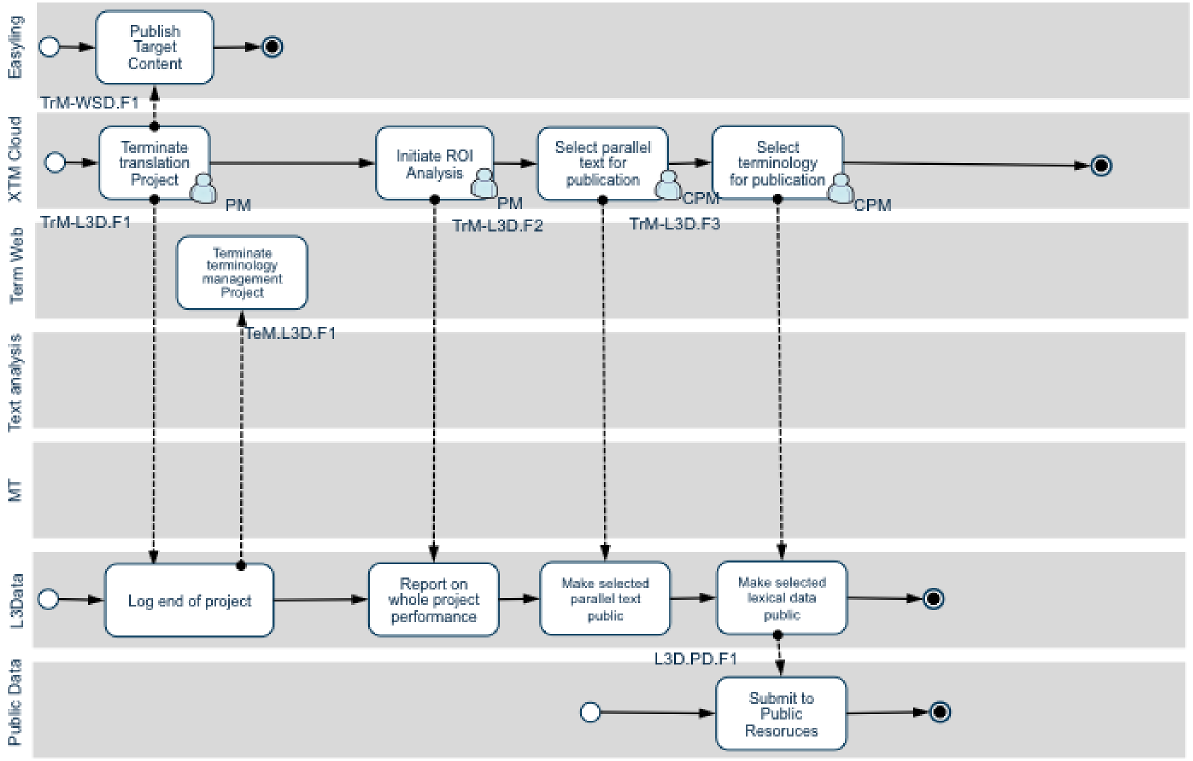

Easyling
XTM Cloud
Term Web
Text analysis
MT
L3Data
Public Data
Publish Target Content
TrM-WSD.F1
Terminate translation Project
PM
Initiate ROI Analysis
PM
Select parallel text for publication
CPM
Select terminology for publication
CPM
TrM-L3D.F1
TrM-L3D.F2
TrM-L3D.F3
Terminate terminology management Project
TeM.L3D.F1
Log end of project
Report on whole project performance
Make selected parallel text public
Make selected lexical data public
L3D.PD.F1
Submit to Public Resources

Figure 10: Sub-process F: Publish Content and Data

ACTIVITY: Terminate Translation Project

- TrM-WSD.F1: Finalise content translation
    - INPUT
        - LSP PM credentials
        - Project name
        - Finalise content in target languages
    - BEHAVIOUR: Publish translated content as target language proxies to source web site.
    - RETURNS
        - URL of published target language proxies
- TrM-L3D.F1: Log end of project
    - INPUT
        - LSP PM credentials
        - Client ID and project name
        - Project data in TIPP file containing XLIFF with final data and post-editor user activity logs.
    - BEHAVIOUR: Log final translations and meta-data annotated as published version
    - TeM.L3D.F1: Log final term base changes
        - INPUT
            - LSP PM credentials
            - Client ID and project dictionary ID
        - BEHAVIOUR: Retrieve final project term base and meta-data annotated as published version
        - RETURNS
            - L3Data reference for log of completed project termbase

  - RETURNS
    - L3Data reference for log of completed project data and meta-data

ACTIVITY: Initiate Project Return-on-Investment (ROI) analysis. LSP PM analyses completed project log to assess the impact of the language technology components used, i.e. automated term extraction and machine translation, and the effectiveness of post-editing prioritisation and iterative retraining of machine translation. The evaluation of automated term extraction is assessed by analysis of the number of suggested terms and the proportion of them that were validated, rejected or that remain un-checked. The number of automated term translation suggestions and the proportion that were validated, rejected or remain uncheck is assessed, broken down by those that were checked in the initial validation check for suggested terms and those term translation that were validated during post-editing. The return on investment is assessed by analysing how frequently term translation validation by these different methods were successfully used in a segment machine translation, i.e. the forced decoding of the term in the MT engine was not subsequently changed in post-editing. If the automated term extraction engine had been used on a previous job for that client then the level of rejected suggestions (i.e. verified false positives) between projects would also be analysed to determine if the term extraction engine successfully improves from the feedback of false positives. The analysis of optimisation of the post-editing is based on the per segment person-time invested in post-editing the output of each iteration of the MT engine and any change in this observed between different iteration, i.e. whether post-editing time per segment (normalised for segment length) has reduced over MT iterations. This may be supplemented by comparing post-hoc automated machine translation quality measures, e.g. translation error rate, for different MT engine instances against a sample of the post-editing output. The time cost to the project will also be analysed, including time spend on term validation, target term capture by post-editors.

- TrM-L3D.F2: Finalise content translation
  - INPUT
    - LSP PM credentials
    - Client ID and project name
    - L3Data filter – specifying which project data is to be retrieved for analysis
  - BEHAVIOUR: Retrieves data from project data set
  - RETURNS
  - L3Data tables covering source text; extracted, validated and translated terms; machine translated and post-edited segment translations; per-segment post-editing timing logs (where permissible); provenance meta-data of these items including reference to engines and human actors involved and processing times.

ACTIVITY: Select parallel text for publication. The client project manager can opt to review the log of parallel text and publish this parallel text under a selected license, specifying if it is free to use and for what purposes, and whether permission or attribution is required, or whether there is payment to use this parallel text. Further, access to parallel text may be restricted to specific agents, e.g. business partners, LSPs, members of specific industry bodies etc.

- TrM-L3D.F3: Select parallel text for publication
  - INPUT
    - Customer PM credentials
    - References to selected parallel data
    - License and access control rule for selected data

- Publication options, i.e. output options, URL configuration, syndication to public aggregators, e.g. DataHub[9], LingHub[10], EU open data portal
- BEHAVIOUR: Publish data as per instructions
- RETURNS
    - URL reference to published data

ACTIVITY: Select terminology for publication. The client project manager can opt to review the log of project terminology and publish parts of it. This can be published in consumer friendly formats such as TBX or the Ontolex vocabulary for lexical semantic resources[11]. They may opt to provide definitions and translation as part of the term base, and may filter which ones are published based on the level and authoritativeness of the validation (e.g. by terminologist, client marketing staff or post-editors with different level of experience). They may opt to filter out term translations that were subject to post-editing (indicating difference in opinions of the correctness of the translation). They may opt to provide examples of use taken from translated project segments, either included in the term base or provided separately as a set of links between the term-base and published parallel text. The publication of links would need to be performed in the context of the agreement with the LSP, since, while the source and target text is typically owned by the client, the LSP may retrain some rights over the links between them. As with parallel text, the client PM can opt to impose different license terms, with different conditions and restrict access to specific parties.

- TrM-L3D.F4: Select terminology for publication
    - INPUT
        - Customer PM credentials
        - References to selected terminological data
        - License and access control rule for selected data
        - Publication options, i.e. output options, URL configuration, syndication to public aggregators, e.g. BabelNet
    - BEHAVIOUR: Publish data as per instructions. Publication to lexical-semantic resource such as BabelNet may benefit from a feedback channel where users of public resources can provide corrections or additions that would be usefully integrated into customer term-base.
    - LD3.PD.F1: Submit term-base to public aggregator
        - INPUT:
            - Term-base to be published
        - BEHAVIOUR: Integrates terminological data into aggregated data set and provide channel to changes and additions related to this data
        - RETURNS:
            - Channel reference for receiving update from aggregator
    - RETURNS
        - URL references to published data
        - Reference to corrections feed from public resources

6. Non Functional Requirements

The following non-functional requirements have also been identified for the FALCON Showcase System:

[9] http://datahub.io/

[10] http://linghub.lider-project.eu/

[11] http://www.w3.org/community/ontolex/wiki/OntoLex_Core_Model

Interoperability Requirements:

- IR1: Interoperability with commercial tools should use data exchange formats that are compatible with established localisation standards as far as possible, e.g. XLIFF for project bi-text, TMX for translation memories and TBX for terminology. This will ease possible future integration of LT and L3Data components with other localisation tools.
- IR2: The definition of open data vocabularies for L3Data should conform to existing and where possible internationally standardised data vocabularies.

Security Requirements:

- SR1: Authorisation of users and exchange of credentials between components should use open mechanisms wherever possible to ease integration of other component in the future.
- SR2: The structure of L3Data should reflect the ownership of different aspects of the data and allow those rights to be unambiguously applied when setting access control mechanisms for different elements of the collected data.
- SR3: Access control mechanisms should be used to ensure the security and integrity of all data.
- SR4: Access control mechanisms should be able to directly reflect: the terms of commercial contracts in place between actors in the value chain, primarily the translation client and the LSP; the terms of applicable data protection legislation, including forthcoming provisions of the EU Generate Data Protection Regulations; and regulation related to the collection worker performance data, e.g. the Works Council Constitution Act (Betriebsverfassungsgesetz - BetrVG) in force in Germany.

Exception Requirements:

Interfaces between components should support reporting of exceptions to ensure robust operation and graceful failure recovery:

- ER1: Authentication Exception: when a user cannot be authenticated.
- ER2: Authorisation Exception: when a user is not authorised to access the requested service or resources.
- ER3: Unknown Resource: when the resource being accessed is unknown to the service.
- ER4: Malformed Data: when data provided is incorrectly structured or formatted.
- ER5: Timeout: when the duration of a session or validity period of an authorisation is exceeded.
- ER6: Server Failure: when some processing or system error occurs and the invoked function cannot complete.

7. Glossary

ATE	Automated Term Extraction
LT	Language technology
L3 Data	Linked Language and Localisation data
LSP	(in this document) Language service provider

LT	Language technology
ML	Multilingual
MT	Machine translation
OntoLex	A lexical semantic linked data vocabulary defined by the W3C OntoLex community group
NER	Named entity recognition
PE	Post-editing (of machine-translated text)
PM	Project Manager
Client PM	Client Project Manager
RDF	Resource Description Framework, a standard model for data interchange on the Web
SMT	Statistical machine translation
TBX	TermBase eXchange standard
TMX	Translation Memory eXchange standard
QA	Quality assurance
XLIFF	XML Localization Interchange File Format