



TaaS

Terminology as a Service

Project no. 296312

Deliverable D3.5 Integrated TaaS system

Version No. 1.0

30/11/2013

Document Information

Deliverable number:	D3.5
Deliverable title:	Integrated TaaS system
Due date of deliverable:	30/11/2013
Actual submission date of deliverable:	30/11/2013
Main Author(s):	Andis Lagzdīns (Tilde)
Participants:	Tilde
Internal reviewer:	IIM
Work package:	WP3
Work package title:	TaaS platform and terminology services
Work package leader:	Tilde
Dissemination Level:	PU
Version:	1.0
Keywords:	Terminology extraction, terminology service, collections

History of Versions

Version	Date	Status	Author (Partner)	Contributions	Description/ Approval Level
0.1	08/11/2013	Draft	Tilde	Document fishbone and input to sections	Drafted by Tilde
1.0	30/11/2013	Final version	Tilde	Updated and finalised document	Revised by Tilde

EXECUTIVE SUMMARY

This is an accompanying document for the deliverable D3.5 Integrated TaaS system, an integral part of the TaaS system.

Table of Contents

Abbreviations.....	4
1. Introduction	5
2. Terminology life cycle	5
3. Architecture overview	6
4. Authentication and authorisation	7
5. Project management	9
5.1. Creating project.....	9
5.2. Project list	10
5.3. Detailed view	10
6. Document upload.....	11
7. Extraction process	12
7.1. Modules	12
7.2. Parameters.....	14
7.3. Status and result	14
8. Terminology management.....	15
9. Terminology sharing	17
10. Terminology reuse.....	17
11. TaaS API and use from Machine Users	18
List of figures	21
List of tables.....	21

Abbreviations

Table 1 Abbreviations

Abbreviation	Term/definition
API	Application programming interface
CAT	Computer-aided translation
DB	Database
DC	Data category
ETB	EuroTermBank
GUI	Graphical user interface
HTML	Hypertext Markup Language
HTTP	HyperText Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IATE	InterActive Terminology for Europe
ISO	International Organisation for Standardisation
JSON	JavaScript Object Notation
MT	Machine translation
REST	Representational State Transfer
SMT	Statistical machine translation
STR	Shared Terminology Repository
TaaS	Terminology as a service
TBX	TermBase eXchange
TDB	Terminology database
TSV	Tab-separated values

1. Introduction

This document accompanies the deliverable D3.5 Integrated TaaS system.

Three application abstraction layers of the TaaS system are described in the Deliverable D3.2 Shared Term Repository.

The integrated TaaS system document analyses the current system and draws conclusions from the following previous deliverables:

- D3.1 Technical requirement analysis and specification
- D3.2 Shared Term Repository – D1.2 analysed user needs and drafted the TaaS platform's functionality in order to satisfy user needs
- D3.3 Facility for term candidates acquisition and term glossary creation
- D3.3 Facility for terminology clean-up and sharing

This document describes the last version of the TaaS system. Some changes have been made since the previous deliverable. After analysing user feedback and to improve the user experience, the TaaS system has dropped the concept of 'user groups', and now users can directly share a project with other users, to whom the user assigns different roles. Also, previously the *project* was called as a *workspace*, but to avoid misunderstanding, the concept was renamed.

2. Terminology life cycle

The terminology life cycle starts with the acquiring and processing of the term collection. The next step is editing and sharing it with others. Finally, the terminology is reused with the various tools – the TaaS web portal, various CAT tools, and in machine translation systems to enhance the quality of translation results (see Figure 1).

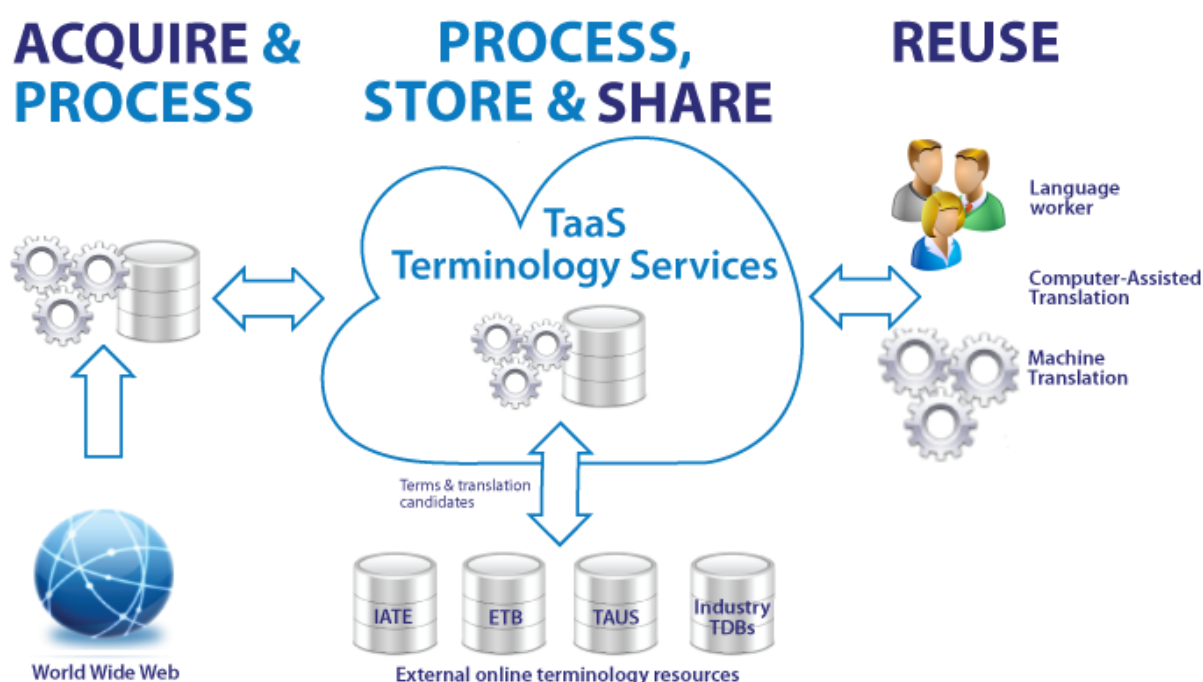


Figure 1 Terminology Life Cycle

The workflow starts with an uploading of the source documents by a TaaS user. The TaaS system identifies term candidates and extracts them into a single collection, as well as finds translation candidates from the Web, industry terminology databases, and other public user collections. After these automatic steps, the user can review the collection and clean it up. The user can share the final collection with others and reuse it in his/her own translation work. Terminology can be used directly from a web interface by search functionality or within the user's CAT tools. Terminology collections can also be used by the supported machine translation systems to increase quality of automatic translation.

The TaaS system provides an iterative terminology life cycle – after one collection is created and shared, the terminology can be used in other terminology extraction projects, where it could be further improved. Thus, terminology is always up-to-date and constantly developing.

3. Architecture overview

The TaaS system has a multi-tier architecture and has three application abstraction layers. It has (i) an interface layer for the user interface (web portal) and APIs with external systems, (ii) an application logic layer for the system logic, and (iii) a data storage layer consisting of file and database storage (see Figure 2).

The interface layer provides interfaces between the TaaS system and end users. The system is used by both human and machine users. Human users access the system through Web browsers by using the TaaS Web page interface. External systems, such as CAT tools and MT users, access the STR services through a public API.

An application logic layer contains a set of modules responsible for the main functionality or logic of the systems. It receives queries and commands from the interface layer and prepares answers or performs tasks using the data storage. This layer contains several modules, such as the Terminology Manager, the User Manager, the Terminology Reuse Manager, the Terminology Extractor, etc. Application logic also includes the terminology extraction service, allowing execution of the term extraction processes asynchronously from the users' requests. The user can start the extraction process and can see the results again later. This approach also allows custom scaling of these resource-intensive modules, that is, separately from the user interface and data layer.

As a data and terminology sharing platform, the TaaS SRT system stores a huge amount of terminology data in various stages - documents, intermediate results, monolingual and bilingual collections. The uploaded documents and intermediate results of term candidate extraction are stored in the file store and registered into the database.

Extracted terminology data is stored in an SQL database in order to be cleaned-up and reused later. Terminology data can be in various stages – extracted monolingual dictionaries, aligned multilingual dictionaries, revised dictionaries ready for terminology reuse.

Data storage consists of three main components:

- STR database – database for the user's acquired terminology collections, his/her profile, group and role in it.
- File store – temporal store for term extraction process and for the files uploaded by the user.
- Web Statistical database – data storage and access components for raw aligned terms.

Detailed description of the TaaS architecture and infrastructure is available in the Deliverable D3.1 Technical requirement analysis and specification.

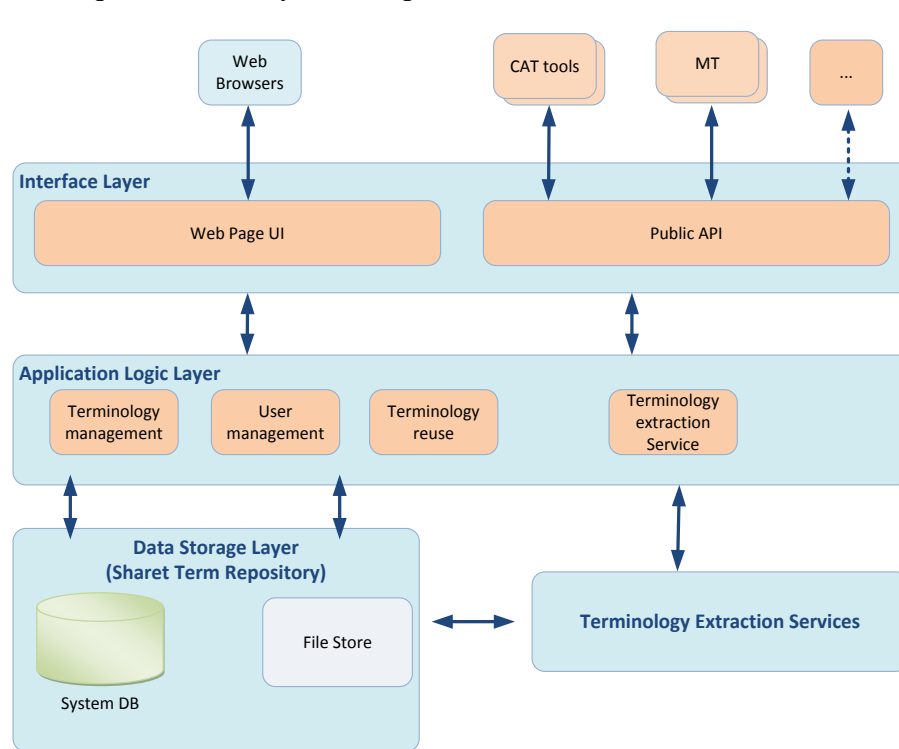


Figure 2 TaaS System Overview

Key components of the TaaS platform are:

- The Web interface for human users (translators, terminologists, etc.).
- The API for machine users (MT systems, CAT tools, etc.).
- The application logic layer.
- The Shared Term Repository (system database for terminology data storage).
- The TaaS platform's File Store (for storage of the users' uploaded document and intermediate results of term extraction).
- Terminology Extraction Services.

The terminology extraction services include various software modules, which are responsible for specific functionality. Key software modules are:

- Modules for extraction of plaintext from user uploaded documents.
- Modules for extraction of term candidates.
- Modules for collection of term translation candidates.
- Modules for retrieval of term translation equivalents.
- Modules for merging extracted term glossaries.

4. Authentication and authorisation

To access the TaaS platform, the user has to provide his credentials. Authentication is implemented via the Tilde ID authentication system (see Figure 3). If the user is new to the

TaaS platform, he can register in the Tilde ID system or log in using external providers such as Google ID or Yahoo ID. The user is added to the TaaS system after the first time he/she has logged into the TaaS system and created a user account. After the first time the user visits the TaaS system, he can fill in basic information about himself, and the initial project is created with administrator privileges in it.

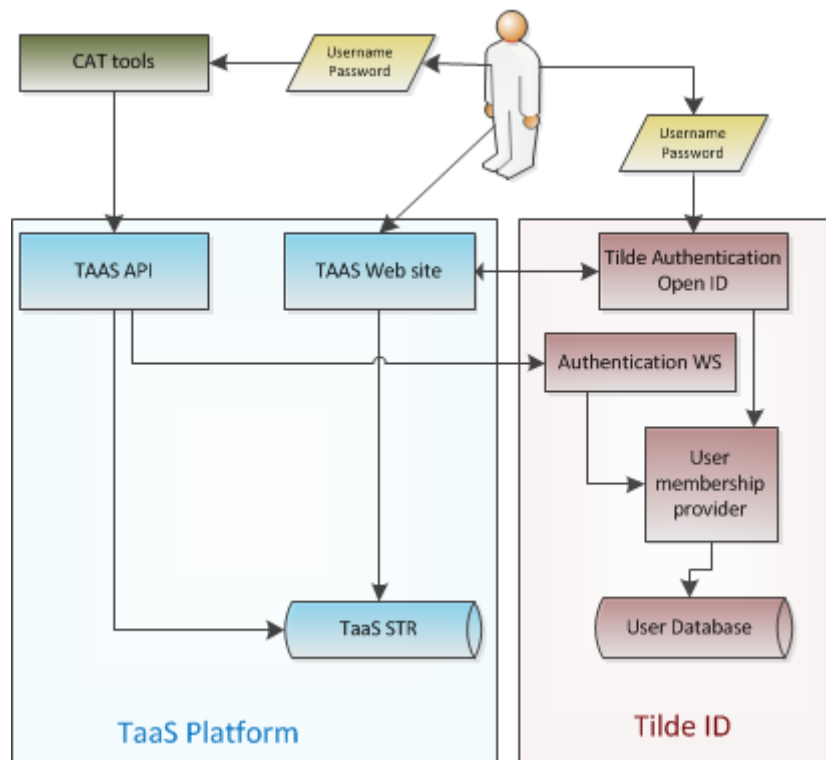


Figure 3 Authentication schema

The TaaS API is accessible using Representational State Transfer (REST) and basic authentication over HTTPS protocol. This means that every request must contain credentials for accessing API methods – username and password encoded with BASE64 encoding schema and include a request in the header.

There are different types of users in the TaaS System:

- TaaS Standard User – can be authorised into the TaaS Web portal and consume the TaaS API.
- TaaS Machine User – can consume the TaaS API and can impersonate a Standard TaaS User in the TaaS API, that is, if the user has approved that the TaaS Machine User has the right to impersonate the user.

To support other services (Machine Users) that act in behaviour of their user access to the TaaS platform, API consumers can add the user key generated by the TaaS system in the request's header. In this way, Machine Users will be able to authorise with their credentials (as TaaS Machine Users) and impersonate a Standard User.

TaaS Machine Users can also perform lookups and execute text extraction methods on public data without impersonating a Standard User, but in that case, there is no access to any private data.

A Machine User account can be created only by request to the TaaS team.

5. Project management

A project is a virtual place where the term extraction process can be executed and the terminology collection managed (see Figure 4). In the initial phase, this place was called a *workspace*, but based on users' feedback, it has been renamed as a *project*. When the user is registered within the TaaS system, an initial project named *Default project* is created. Every project has its owner – a user who has created it. He can invite other users to collaborate in that project within different roles. Every user can own and share multiple projects.

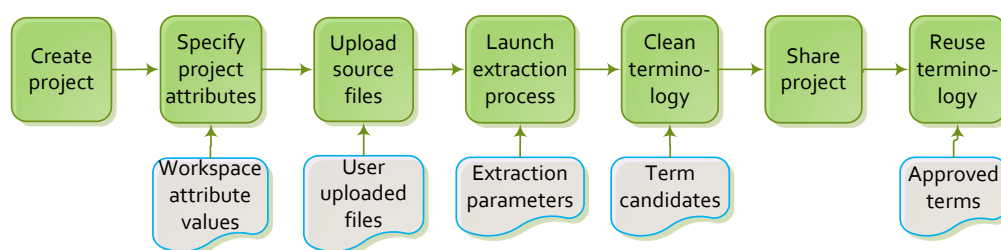


Figure 4 Terminology workflow within a project

5.1. Creating project

The lifecycle of a project starts with its creation (see Figure 5). After that the user has to fill out the form by specifying workspace attributes. It is important to enter these attribute values before the extraction process because the extraction process will extract terms based on these attributes. Also, the created term collection will inherit these attribute values.

Create Project

Name <input type="text" value="Terminology project"/> *		Description <input type="text" value="Project for specific purposes. Very confidential!!!"/>	
Source Language <input type="text" value="English"/> *	Target Language <input type="text" value="Latvian"/> *	Domain <input type="text" value="Industries and technology"/> *	
Product <input type="text" value="Specific product"/>	Customer <input type="text" value="Important customer"/>	Business Unit <input type="text" value="Business Unit"/>	

☐ Public collection (available for search and lookup by other TaaS users)
☒ Private collection (available for search and lookup by project users)

Figure 5 Creation of Project

The following attribute values are mandatory for the terminology extraction process:

- Name – name of the project and extracted collection in plain text, cannot be longer than 200 characters.
- Source language – one of the TaaS system's supported languages (25 supported languages).
- Target language – one of the TaaS system's supported languages (25 supported languages).
- Subject field, or domain, – one of TaaS system's supported subject fields. This attribute is mandatory.

There are some attributes that are optional:

- Description – description of the project and extracted collection in plain text, cannot be longer than 1000 characters.
- Product – information to indicate that it is related to or used in a specific product. Plain text, cannot be longer than 100 characters.
- Customer – information associated with a specific customer. Plain text, cannot be longer than 100 characters.
- Business Unit – information associated with a specific department, division, or other unit of an enterprise. Plain text, cannot be longer than 100 characters.

Once a project has been created, the user comes to the place for project management where the user can proceed to work with terminology acquisition. Also, this newly created project appears in the common project list.

5.2. Project list

The project list is a place where the user can see all his owned projects and the projects that are shared with other users. The project list has the following columns:

- Name – project name. If the project is shared, a sharing symbol appears next to the name.
- Status – whether the project is made public or not.
- Source – the source language of the project.
- Target – the target language of the project.
- Domain – domain of the project.
- Owner – the user, who is the owner of the project.
- Created – date of project creation.
- Updated – last date of project update.

All columns can be sorted as ascending or descending, the default is name ascending.

5.3. Detailed view

Once a project has been created, the user comes to the place for project management where the user can proceed to work with terminology acquisition. Also, this newly created project appears in the common project list.

The detailed page of the project consists of four sections:

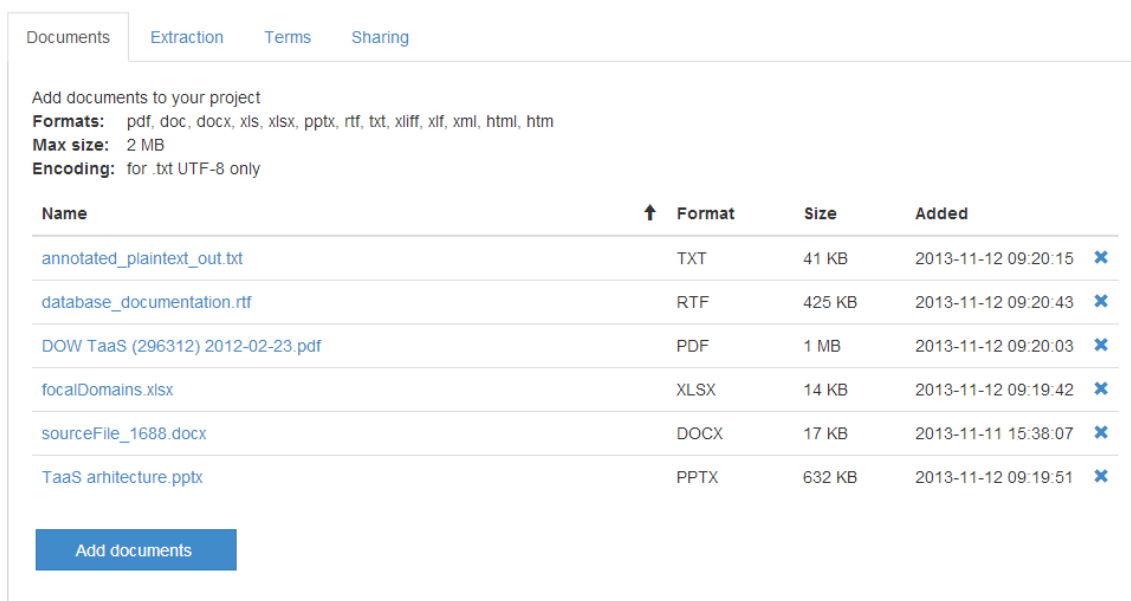
- Documents – source document management.
- Extraction – extraction parameters and status.
- Terms – terminology management.
- Sharing – list of users with whom this project is shared. This is also where additional users can be added to the sharing list.

From the detailed page of the project, the project can be deleted and the project metadata can be edited. By deleting the project, files uploaded by the user and the related terminology collection will also be deleted.

6. Document upload

Uploading files to the project is a prerequisite for launching the terminology extraction process. At least one file must be uploaded before the workspace is prepared and ready for the extraction process.

The user has the possibility to upload multiple files, with *Drag-and-Drop* also as a feature. When the user has uploaded files he/she can see the list of files, their format, size, and upload date, and can delete or download them to his/her computer locally. The list of documents can be sorted by any column – *Name*, *Format*, *Size*, or *Added* (see Figure 6).









Name	↑ Format	Size	Added
annotated_plaintext_out.txt	TXT	41 KB	2013-11-12 09:20:15 
database_documentation.rtf	RTF	425 KB	2013-11-12 09:20:43 
DOW TaaS (296312) 2012-02-23.pdf	PDF	1 MB	2013-11-12 09:20:03 
focalDomains.xlsx	XLSX	14 KB	2013-11-12 09:19:42 
sourceFile_1688.docx	DOCX	17 KB	2013-11-11 15:38:07 
TaaS architecture.pptx	PPTX	632 KB	2013-11-12 09:19:51 

Figure 6 Document Section of Project

The maximum size of every document that can be uploaded is 2 MB, and the following file formats are supported for upload – pdf, doc, docx, xls, pptx, rtf, txt, xliff, xlf, xml, html, and htm.

Files are uploaded to the TaaS File Store, where every user has their own file directory. Also, a directory, where the user's uploaded files are stored, is created for every project.

Terminology extraction services access this File Store and extract information from the uploaded files.

7. Extraction process

When a project has at least one document added and the project's properties are set, the user can trigger the execution process – the TaaS system takes the attributes entered by the user, the list of uploaded files, and the workflow configuration, creates a definition file for terminology extraction, and starts the process.

The execution of terminology extraction runs asynchronously from the user graphical interface, so that the user can return later to see the results of terminology extraction. During the terminology extraction process, the user has the possibility to follow the status of the extraction process and also to review intermediate results. As the result of the terminology extraction process, there is an extracted terminology collection.

7.1. Modules

The extraction process consists of several independent software modules that are organised in a workflow (see Figure 7). The extraction workflow consists of the following steps:

- Plain text extraction.
- Monolingual term candidate acquisition.
- Translation candidate lookup.
- Result processing.

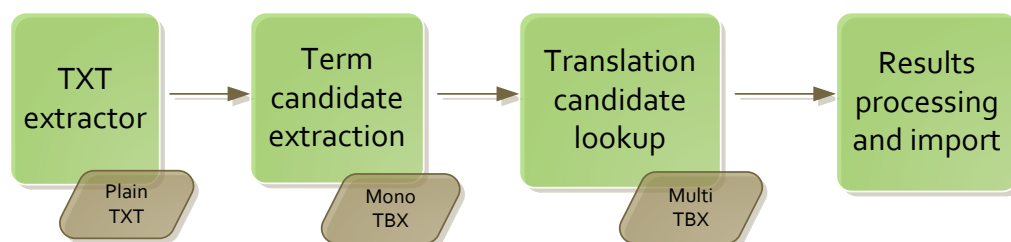


Figure 7 Terminology Extraction Workflow

Plain text extraction

The first step of the terminology extraction process is the plain text extraction from the user's uploaded document. As the user can upload different types of files in different encodings, this module is responsible for the correct interpretation of a file structure, encoding, and content. At first, the main module of plain text detects file format and encoding, takes the user entered parameters and decides which tool will be used for text extraction. The plain text extraction process is executed for each file, and the result is the TXT file that is saved in the TaaS File Store. This text is an intermediate result and is used later as an input for the later software modules in the workflow.

Term candidate extraction

Monolingual term candidate extraction includes a set of software modules that are responsible for the acquisition of monolingual term candidates in the source language selected by the user. As input, these modules receive the plain text from the source documents and the project

properties provided by the user. The result after these modules is a monolingual TBX document. Source term candidate extraction modules include (see Figure 8):

- *Term markers in plain text* – currently there are two independent modules which can be selected by the user before the extraction process – TWSC tool and Kilgray Term Extractor. If both tools are selected, these tools are executed in parallel, and the results are later merged (full merge).
- *Term normaliser* – appends canonical forms and morpho-syntactic information about canonical forms to the marked terms of documents.
- *Single Collection Creator* – compounds all terms from tagged documents into a single terminology collection. The terms are grouped by canonical forms and morpho-syntactic information, and a single monolingual terminology collection is created including references to the source files and collocations.

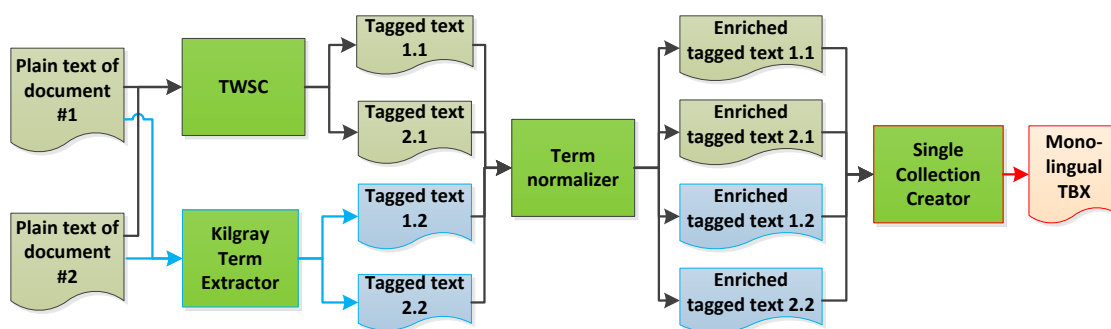


Figure 8 Monolingual Term Extraction

Translation candidate lookup

After the monolingual terminology collection is created, lookup for translation candidates in the target language provided by the user is performed. Lookup for translation candidates occurs as separate processes which are executed in parallel to every source selected by the user.

The most reliable results are acquired from externally linked terminology databases – EuroTermBank, IATE, or TAUS. Lookup is also performed in STR – in public user collections and in the user’s private and shared collections. Lookup in the Web Statistical Database is less reliable.

The result after lookup is one multilingual TBX for each lookup resource, with translation from the corresponding terminology source.

Result processing and import

After the translation lookup in the target language, all of the bilingual terminology collection versions are merged into a single glossary, creating one TBX file. Terms are grouped by source language terms, and duplicates in the target language elements are removed.

The last step of terminology extraction is to import the acquired terminology collection into the TaaS database. This is done using the public TaaS API, during which time, the extraction service impersonates the user to perform the import process.

7.2. Parameters

The extraction process is configurable. Configuration is divided into three blocks:

- Document selection – the user can select all or specific documents to be used in extraction. One document must be selected to start the extraction process.
- Term extraction – the user can select tools that will be used in the source term extraction (at minimum, the *TWSC* or *Kilgray* tool must be selected):
 1. *TWSC* - [Tilde wrapper system for CollTerm](#) is based on linguistic analysis (part of speech tagging, morpho-syntactic patterns, etc.) enriched by statistical features (e.g., frequency score).
 2. *Kilgray* - [Kilgray Terminology extractor](#) is based on language independent statistical analysis (e.g., frequency score).
 3. *Normalise terms* - Canonical, or dictionary, forms (recommended).
- Translation lookup – possibility to select translation lookup sources in the target language (at least one source must be selected):
 1. *TaaS public collections* - collections shared by other TaaS users, making them publicly available.
 2. *My collections* – Collections owned by the user or collections belonging to other users that are shared with this user.
 3. *EuroTermBank* – [EuroTermBank](#) (ETB) is the largest European online term bank that stores more than 2 million terms in 27 languages.
 4. *Inter-Active Terminology for Europe* - [Inter-Active Terminology for Europe](#) (IATE) is an inter-institutional terminology database of the European Union. This might take a few more minutes.
 5. *TAUS Data* - [TAUS Data platform](#) stores shared translation memories. This might take a few more minutes.
 6. *Web Data* - TaaS Statistical Database (SDB) stores data automatically extracted from the Web.

7.3. Status and result

During the extraction process, the user can follow its status – see the current execution step. There is no need to wait on that page till the end of processes, the user can work with other projects or come back to the system later to see the results. The duration of the extraction process depends on:

- User uploaded documents – number of documents and their total size and complexity.
- Project properties – for some languages or domains, data analysis is more complex and therefore consumes more resources and time.
- Selected tools – some tools may take more time to execute.
- Queue length waiting on extraction service – all user extraction processes are queued and wait for free resources. There may be a situation when a large number of users starts the extraction process at the same time.

The user can stop the extraction process at any time by reallocating the resources on the server (see Figure 9). During the extraction, no project property changes are allowed, and therefore, the user might want to stop the process in order to change the settings.

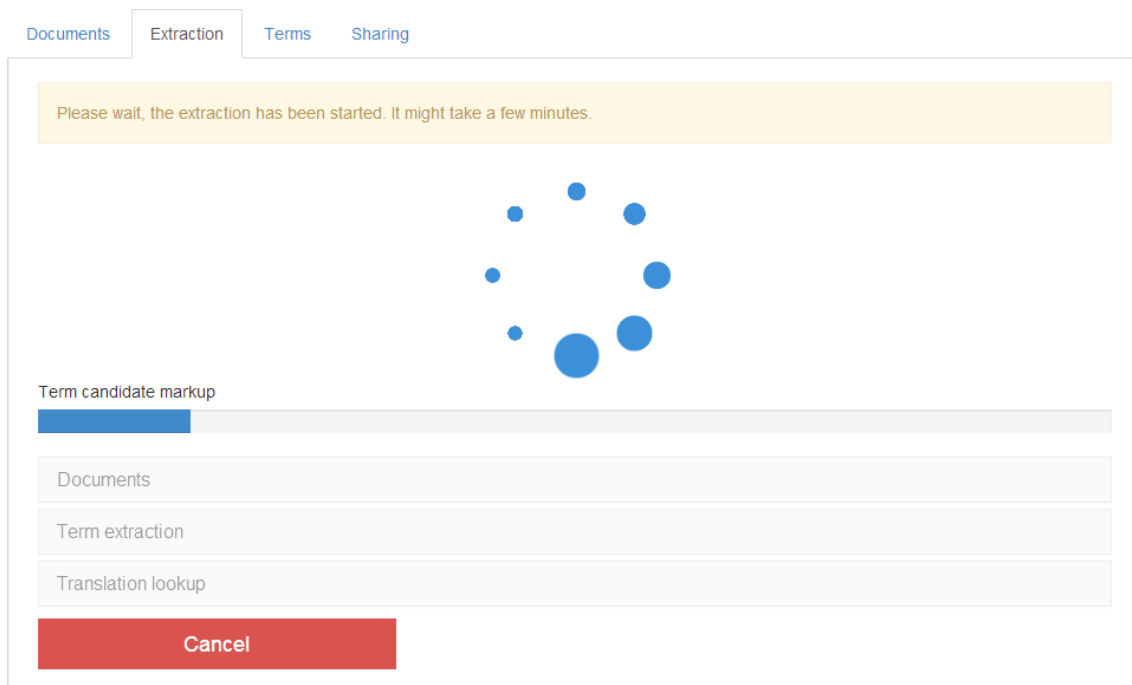


Figure 9 Extraction in Progress

The result of extraction is a terminology collection. The terminology collection appears in the *Terms* section of detailed view of the project.

8. Terminology management

Visual representation of acquired terms and their translation is available immediately after the extraction process has completed. The list of terms is ordered by the amount of translations and candidates in descending order– first come the terms with more translations and translation candidates.

Each line represents a term entry, and the following columns are visible:

- Extracted terms for the source language.
- Approved term translations and the feature to *add new translation* by typing in the text box. With a single click on a translation, the translation can be removed from the list.
- Translation candidates. Translation candidates are grouped by sources (visualisation in different colours), and only the top three translations per source are visible; other translations can be loaded by action of an additional user.
- Term entry operations – image buttons for advanced editing and deletion.

List of term entries are paged – every page contains 30 terminology entries. At the top of each list of terms, the user has these filtering possibilities (see Figure10):

- Filter out by the first character of the source terms.
- Filter by search text – selects terms that includes the entered text.
- Hide terms not approved.
- Hide terms without translations.
- Choose sources from which translation candidates are provided.

Documents	Extraction	Terms	Sharing
All	Other	A	Ā
B	C	Č	D
E	Ē	F	G
Ģ	H	I	Ī
J	K	Ķ	L
Ļ	M	N	Ņ
O	P	Q	R
S	Š	T	U
Ū	V	Z	
Search in extracted terms			
Extracted terms	Approved translations	Translation candidates	Import
maksa	add translation	<div>ETB</div> <div>Web</div> charge payment maksa fees charge +6 more	✎ ✕
rēķins	add translation	<div>Web</div> <div>TaaS</div> expenses invoice accounts +7 more account	✎ ✕
laiks	time add translation	<div>ETB</div> <div>Web</div> time laika historic time time +5 more	✎ ✕
konference	add translation	<div>Web</div> konferencei Latvia conference conference +6 more	✎ ✕
personāls	personnel staff personnel add translation	<div>ETB</div> <div>Web</div> personnel staff staff personnel personal +3 more	✎ ✕
reģistrācija	registration add translation	<div>ETB</div> <div>Web</div> REG registration registration registrācija means registration +2 more	✎ ✕

Figure 10 List of Extracted Terms

Initially the terminology collection has no approved terms, and the TaaS system provides the term candidates for the source language and the translation candidates for the target language. The user revises collections by accepting good terms, editing them, and deleting unnecessary terms, as necessary. The terms can be shared for use in other data reuse scenarios only after the approval process.

There are two data revising scenarios implemented in the TaaS web portal:

- **Fast term and translation candidate approval** with restricted functionality.
- **Advanced term editor:** After the approval process, the user can chose advanced editorial capabilities – term editing, adding definitions, explanations, references, linguistic, and other information.

The term representation view also provides functionality for importing and exporting terminology collections in the following formats:

- TBX (TermBase eXchange format).
- TSV (Tab-separated values).
- CSV(Comma-separated values).

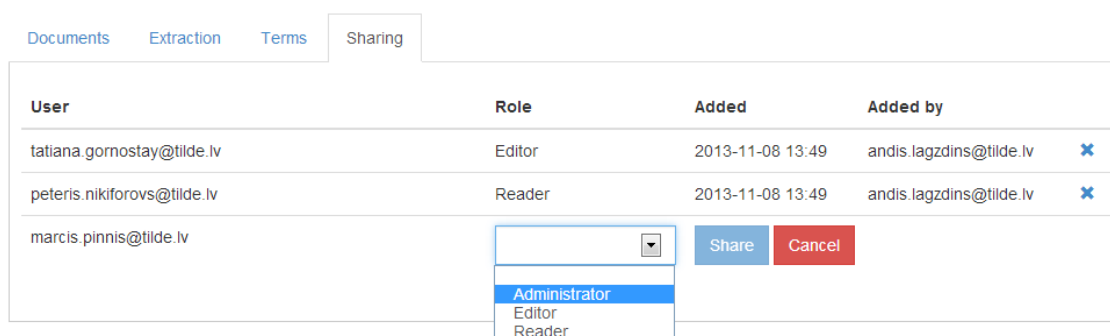
9. Terminology sharing

After the user has cleaned and revised his terminology collection, the user can now reuse his created terminology. To provide the possibility for other users to use the revised terminology, there are the following possibilities:

- Switching the project's terminology collection status from private to public.
- Giving specific users access to the project in the *Sharing* section.
- Downloading the entire collection in the TaaS provided data format and sharing it independently from the TaaS system.

In the section named *Sharing*, the user can choose to share the project with others. A project can only be shared with users who are already registered in the TaaS system by providing their email addresses. When sharing, the user has to provide the role of the target users; every role defines a different set of permissions for access to the project data and operations. There are three user roles in the TaaS System (see Figure 11):

- Administrator (full-access).
- Editor (limited access).
- Reader (limited access).



User	Role	Added	Added by
tatiana.gornostay@tilde.lv	Editor	2013-11-08 13:49	andis.lagzdins@tilde.lv
peteris.nikiforovs@tilde.lv	Reader	2013-11-08 13:49	andis.lagzdins@tilde.lv
marcis.pinnis@tilde.lv	<div>Administrator Editor Reader</div>		

Figure 11 Project Sharing

10. Terminology reuse

Shared data can be used for various purposes in the following ways:

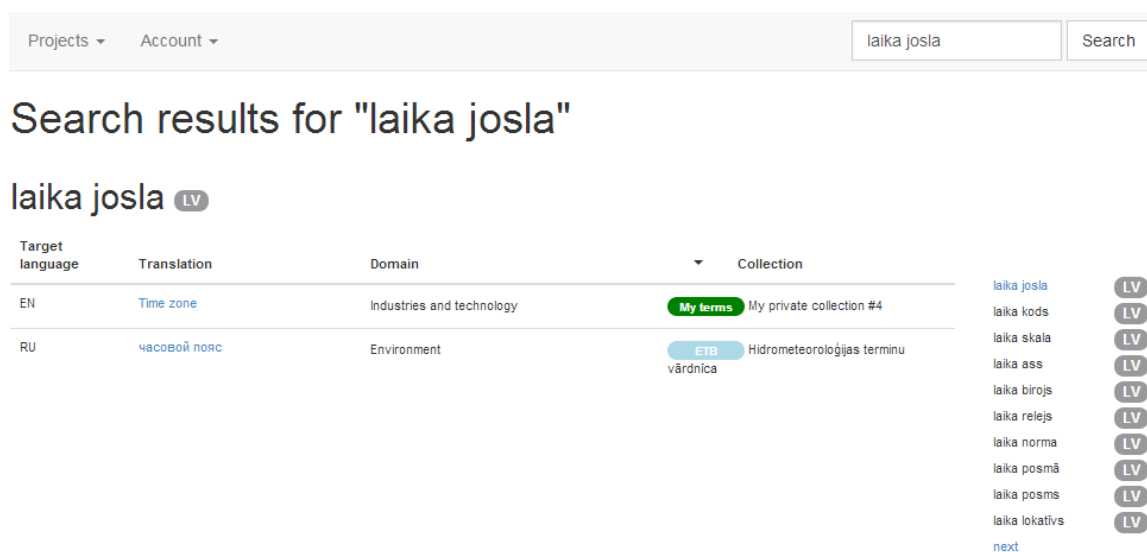
- In terminology searches.
- In other terminology extractions – provide translation candidates for new collections.
- Users can reuse their terminology collections in various CAT tools scenarios.
- Users can reuse their terminology collections in various MT scenarios.
- By downloading the whole terminology collection to be used in other systems.

Users with appropriate permissions can share their collection by setting its type as public, and so, other users will be able to reuse it. However, other users cannot edit or take ownership over such public collections.

The TaaS system provides search functionality both via web portal and TaaS API. The TaaS web portal also provides search functionality for anonymous users – however, searches will only be performed on public collections.

A revised terminology collection can also be used in the other projects. Before extraction, the user can choose translation sources. In case the user has left the source *TaaS public collection* checked, during the extraction, the translation lookup module will search for translation candidates in collections that are marked as public.

Translation lookup source *My collections* means that, during the extraction, translation lookup will search in the private collections owned by the user and also in collections that are shared and to which the user has permissions (see Figure 12).



The screenshot shows the TaaS web portal search results for the query "laika josla". At the top, there are navigation links for "Projects" and "Account", and a search bar containing the query "laika josla" with a "Search" button. Below the search bar, the title "Search results for 'laika josla'" is displayed. The main content area shows the search results for "laika josla" in Latvian (LV). A table lists the results with columns for Target language, Translation, Domain, and Collection. The first result is for EN (English) with the translation "Time zone", domain "Industries and technology", and collection "My private collection #4". The second result is for RU (Russian) with the translation "часовой пояс", domain "Environment", and collection "Hidrometeoroloģijas terminu vārdnīca". To the right of the table, there is a list of related terms: "laika josla", "laika kods", "laika skala", "laika ass", "laika birojs", "laika relejs", "laika norma", "laika posmā", "laika posms", and "laika lokatīvs", each with a small "LV" icon next to it. A "next" link is also visible at the bottom right of the list.

Target language	Translation	Domain	Collection
EN	Time zone	Industries and technology	My private collection #4
RU	часовой пояс	Environment	Hidrometeoroloģijas terminu vārdnīca

- laika josla LV
- laika kods LV
- laika skala LV
- laika ass LV
- laika birojs LV
- laika relejs LV
- laika norma LV
- laika posmā LV
- laika posms LV
- laika lokatīvs LV

next

Figure 12 Search Results

11. TaaS API and use from Machine Users

A user can also access his/her data via CAT tools, MT systems, or other terminology databases (see Figure 13). The TaaS API is accessible using Representational State Transfer (REST) and provides methods for accessing TaaS public and private data and terminology extraction methods.

Machine Users can query data with or without impersonating the end user via TaaS API. To support other services (Machine Users) that act in behaviour of their user to access the TaaS platform, API consumers can add the user key generated by the TaaS system in the request's header. In this way, Machine Users will be able to authorise with their credentials (as TaaS Machine Users) and impersonate a Standard User by providing the User Key. The TaaS Machine User can also perform lookups and execute text extraction methods on public data without impersonating a Standard User, but in that case, there is no access to any private data.

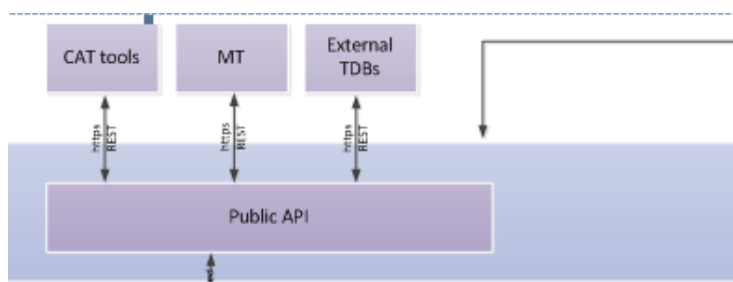


Figure 13 TaaS API for Machine Users

A user can decide whether or not to allow a specific Machine User to act in behaviour of him. Issued API Keys can be managed in the account section named *API Key*. Here the user can enable/disable issued keys or delete them. A user can only generate new API keys together with the Machine client request.

Following services are available via API:

- **Term lookup** (translation) – this method looks up translations for a given term and optionally performs look-ups in the specified target language and domain. If the authorised user is a Machine User account without impersonation, the lookup is performed only in public data.
- **Term Extraction** – this method tags term candidates in text: method 1 marks only term candidates, and method 2 also includes references to terminology entries, i.e., marks term candidates and also searches in TaaS external resources (e.g., Eurotermbank.com) for translations. If the authorised user is a Machine User account without impersonation, lookup is performed only in public data.
- **Retrieve Entry** – this method retrieves an entry by its entryID. If the collection is private and not shared, only a user authorised by the owner can access the entry. If the authorised user is a Machine User account without impersonation, only a public entry can be retrieved.
- **Insert a New Entry or Update an Existing Entry** – this method updates an existing entry or inserts a new entry. If an entry contains the attribute “id” at entry level in the TBX document, the collection ID is retrieved from DB. If the entry id is not present, the API checks the received collection id. If the user has appropriate rights (editor or administrator of this collection), the entry is inserted as a new entry in this collection. If the entry is a new version of an existing entry, the existing entry is replaced with the updated version. After insertion, all search tables are immediately updated. If the authorised user is a Machine User account without impersonation, no action is performed.
- **Get a List of Collections** – this method retrieves a list of available collections. If a filter is passed along the request, the collections are filtered using the filter parameters. Available filters are: a list of languages and a list of domains. If the authorised user is a Machine User account without impersonation, only public collections are returned.

- **Get a Collection** – this method retrieves all entries of a selected collection. If the authorised user is a Machine User account without impersonation, only entries from public collections are returned.
- **Get a List of Domains** – this method retrieves a list of available domains from the TaaS domain classification;
- **Get a List of Languages** – this method retrieves a list of languages supported by the TaaS platform.

List of figures

Figure 1 Terminology Life Cycle.....	5
Figure 2 TaaS System Overview.....	7
Figure 3 Authentication schema.....	8
Figure 4 Terminology workflow within a project.....	9
Figure 5 Creation of Project	9
Figure 6 Document Section of Project	11
Figure 7 Terminology Extraction Workflow	12
Figure 8 Monolingual Term Extraction	13
Figure 9 Extraction in Progress	15
Figure 10 List of Extracted Terms	16
Figure 11 Project Sharing.....	17
Figure 12 Search Results.....	18
Figure 13 TaaS API for Machine Users	19

List of tables

Table 1 Abbreviations	4
-----------------------------	---