TaaS

Terminology as a Service

Final Publishable Summary

Grant Agreement number: 296312
Funding Scheme: FP7-ICT-2011-SME-DCL
Period covered: From June 2012 to May 2014
Project coordinator: Dr. Andrejs Vasiļjevs, Tilde
Tel: +371 76705001
Fax: +371 67605750
E-mail: andrejs@tilde.lv
Project website: www.taas-project.eu

Table of contents

1. Executive summary

The core objective of the TaaS project is to align the speed of terminology resource acquisition with the speed at which content is created by mining new terms directly from the Web. The motivation of the TaaS project is to facilitate terminology work in practical usage scenarios by providing a number of online terminology services.

TaaS has embraced the power of cloud computing technology and latest advances in multilingual data processing by developing a collaborative cloud-based platform for multilingual terminology services. TaaS addresses the needs of language workers by providing online terminology services for key terminology tasks – identification of term candidates in users' documents, lookup of translation equivalent candidates in existing terminology resources, acquisition of translation equivalent candidates from parallel and comparable corpora acquired from the Web, collaborative working environment, crowd-sourced refinement and approval of term candidates and their translations, sharing terminology with other users, and its usage in other working environment – thus becoming a part of a multifaceted global cloud-based service infrastructure.

TaaS has demonstrated its services in three usage scenarios: terminology work by a language worker, computer-assisted translation (CAT), and machine translation (MT). For language workers, the TaaS platform simplifies the management of task-specific multilingual terminology, including its processing, storage, sharing, and reuse via rich functionalities of the TaaS Web-based Graphical User Interface (GUI). For CAT, TaaS provides an instant access to terminological data and services via its Application Program Interface (API) integration with popular CAT tools memoQ and OmegaT. For MT, TaaS boosts translation quality and facilitates domain adaptation by proposing novel methods of terminology integration in MT systems via its API. TaaS evaluation results show that terminological adaption of translation and language models and dynamic pre-processing of translatable content in statistical machine translation (SMT) lead to a significant boost in translation quality by up to 26.9%. Thus, the TaaS API enables the integration of terminology services in a wide range of applications and solutions. For example, the TaaS services are used in the Web-based showcase that demonstrates the application of the Internationalisation Tag Set (ITS) standard (in its updated 2.0 version and recommended by the WWW Consortium) for enriched terminology annotation.

TaaS has attracted significant interest from user communities. During the first weeks of the TaaS open Beta with fully operated services, more than 800 terminology projects were created by more than 750 registered users. TaaS is widely endorsed via social and professional networks Facebook, Twitter (the TaaS account with more than 220 followers), LinkedIn (the TaaS group with more than 730 participants), Google+, ResearchGate, and Scoop.it!. TaaS was presented at 39 academia and business events and its results were described in 25 publications.

The project has fully achieved its objectives by providing innovative terminology services that can become an essential part of the European multilingual infrastructure to serve the critical needs of the public sector and businesses for enabling language technologies. TaaS has a great potential in its integration with machine users for multilingual digital content processing, for example, authoring, translation, indexing, search, and others.

2. Context and objectives of the project

Consistent, harmonised, and easily accessible terminology is an extremely important prerequisite for unambiguous multilingual communication in EU and global world. Every day the volume of terminology is growing along with the explosion of information available on the Web. Current static models for the acquisition, sharing, and using terminological data cannot keep up with the growing demand. Surveys have shown that translators, editors, technical writers, and other language workers spend up to 30% of their working time on terminology research looking for terms in multiple local and online resources, acquiring terminology, and organising proprietary terminology. The role of terminology is more important than ever in the multilingual Europe to insure that people communicate efficiently and precisely. Efficient terminology acquisition and management has become an essential component of intelligible translation, localisation, technical writing, and other professional language work.

Terminology is among the most important language resources providing lexical designations assigned to concepts, term equivalents in different languages, their definitions, usage contexts, and other data. Digital representation of terminological data is in everyday use by language workers and machine users, for example, machine translation, information extraction, semantic search, and other applications.

A typical practice in preparing terminology for translation and writing usually involves manual or semi-automated analysis of documents to identify candidate terms, which are then checked in existing terminology resources for corresponding entries to create a new terminology resource. A conventional model for the creation of a bi-(multi-)lingual terminology resource in this scenario involves (1) the collection of domain specific documents, (2) term identification and preparation of terminology, (3) lookup for matching terminological data in prescriptive terminology resources, (4) creation of new terminological data for entries that are not found in other sources. This work is usually carried out by an individual expert or a group, involves a great deal of manual work, and is rarely shared.

Recently, significant progress has been achieved in the automation of terminology tasks. Web crawling tools have been successfully adapted and applied to collect corpora for terminology needs on the Web. Although several tools are available to automate the individual steps of terminology work, there is still no solution that would cover all key tasks of terminology work. In addition, huge efforts have been invested to consolidate and harmonise terminology resources in national and international online term banks (for example, Rikstermbanken, InterActive Terminology for Europe (IATE), EuroTermBank (ETB)) and domain-specific online dictionaries (for example, Online Nautical Dictionaries, Medical Dictionary, Glossary of Translation and Interpreting Terminology and others). However, these resources are still limited in their coverage and they struggle to cope with a constant need to incorporate an increasing number of new terms resulting from the dynamic pace of technological, scientific, and social changes.

New terms are coined every day by businesses, translation and localisation agencies, collective and individual authors. Although these terms can be found in different online and offline publications, the inclusion of new terms in public terminology databases takes months or even years, if happens at all. For this reason, terminology databases fail to provide users with extensive up-to-date multilingual terminology, especially for terms in

under-resourced languages or specific domains that are poorly represented in online terminology databases. As a result, the major deficiencies of existing terminology resources are high cost and time needed for their creation, insufficient coverage of languages and domains, particularly for the most recent concepts, insufficient sharing of terminology resources, and lack of collaborative mechanisms for the involvement of terminology practitioners. Due to laborious manual work and the incompleteness of terminological data, it is still very time consuming to find and prepare terminological data needed in practical translation work. Several surveys show that technical translators spend more than 30% of translation time on terminology work.

With the evolution of the Internet and cloud computing, there has been a shift towards collaborative solutions with a focus on user and consumer-orientation, consistency, interoperability, and sharing in information management and professional communication communities. Effective processing and use of terminology is the backbone behind robust processes within the content life cycle from its creation, translation, localisation, publication, and numerous other information management steps to ensure efficient and precise communication.

The motivation for the TaaS project is to address the need for instant access to the most up-to-date terms, user participation in the acquisition and sharing of multilingual terminological data, and efficient solutions for the reuse of terminology resources. The objective of the TaaS project is to embrace terminology in the cloud service paradigm by establishing an innovative sustainable online platform that provides terminology services for key terminology tasks. TaaS ambition is to exploit the cutting-edge research advancements to automate terminology work for all official EU working languages – identification of term candidates in users' documents, lookup of translation equivalent candidates in existing terminology resources, acquisition of translation equivalent candidates from parallel and comparable corpora acquired from the Web, collaborative working environment, crowd-sourced refinement and approval of term candidates and their translations, sharing terminology with other users, and its usage in other working environment – thus becoming a part of a multifaceted global cloud-based service infrastructure.

3. Main results of the project

The project has fully achieved its objectives by establishing a cloud-based platform by providing **innovative multilingual terminology services** that cover key tasks of terminology work (see Figure 1):

- **Search** and consolidated **representation** of terms from various sources: TaaS Shared Term Repository and online terminology resources (for example, EuroTermBank, IATE).
- **Identification** and **extraction** of monolingual term candidates from documents uploaded by users using the latest advances in linguistically and statistically motivated terminology extraction.
- **Lookup** of term translation equivalent candidates (for monolingual term candidates automatically extracted from documents uploaded by users) in the largest publicly available terminology databases (for example, IATE, EuroTermBank, TAUS Data) as well as statistical terminological data acquired from the Web.

- **Acquisition** of statistical terminological data from parallel and comparable corpora acquired from the Web using cutting-edge techniques for linguistically and statistically motivated terminology extraction and bilingual terminology alignment.
- **Creation** of monolingual and bilingual terminology collections in user-defined languages in 25 project languages.
- **Collaborative** terminology **refinement** and **approval** of raw terminology extracted automatically, for example, deletion of irrelevant or unreliable term candidates and "incorrect" extraction (for example, a part of a longer noun group or irrelevant terms); definition of termhood and unithood; term variant identification; deduplication; bilingual checking of translation equivalents and deletion of irrelevant or unreliable translation equivalents; validation term candidates in context etc.
- **Sharing** of terminology projects with other users and **sharing** of terminology with external terminology resources.
- **Reuse** of terminology collections in various applications within different human and machine usage scenarios via the TaaS API and export of files in different formats widely exploited by users, for example, TSV, CSV, TBX, custom tabular format, and the Moses SMT system-compliant export format.

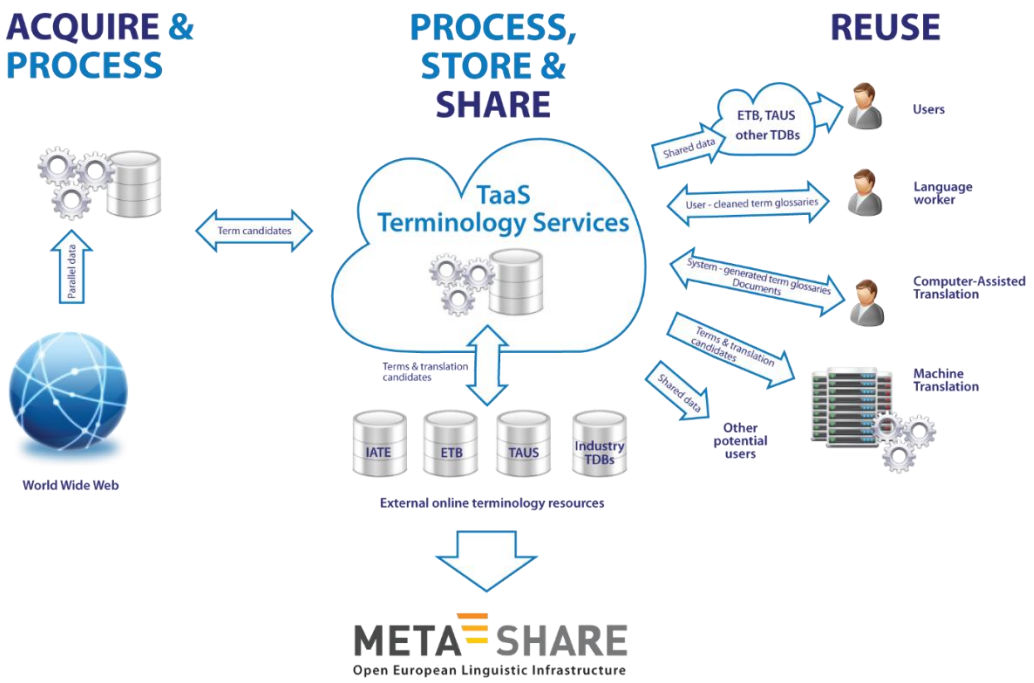


Figure 1. TaaS concept

To fulfil its objectives, TaaS addressed a wide range of challenging scientific and technological tasks.

A detailed **analysis of user needs and requirements** for terminology services was conducted as an online survey with about 1,800 respondents.

Monolingual term candidate extraction was researched, monolingual term extraction modules were developed and integrated in the platform for all official EU working languages and Russian. **Gold Standard** guidelines and evaluation sets were developed for the evaluation of monolingual term extraction and automatic and manual techniques were used for the evaluation of bilingual term alignment.

Tools for **parallel and comparable data acquisition from the Web** were adapted, elaborated, and integrated within the TaaS Bilingual Terminology Extraction System (BiTES). BiTES includes **four workflows integrating extraction and alignment modules from parallel and comparable data** such as multilingual websites, Wikipedia, and online newsfeeds. Parallel and comparable data for terminology extraction were acquired from the Web and processed for the extraction of term candidates.

The TaaS platform uses **scalable layered architecture** to provide robustness and efficiently cope with increasing workload.

User-friendly online interface was created to support typical terminology work scenarios. Particular attention was paid to ensure the ease of use with rich functionality and customisation options.

To provide terminology support in the translator working environment, the TaaS terminology services were integrated in the production versions of **popular CAT tools** memoQ and OmegaT. The **TaaS API** is provided for developers to facilitate integration in other tools.

Several novel methods were elaborated for **improving the quality of statistical MT** by imposing domain-specific terminology. Evaluation results show that terminological adaption of translation and language models and the dynamic pre-processing of translatable content lead to significant boost in translation quality by up to 26.9%.

One of the outstanding results of the project is the integration of the TaaS terminology services with the ITS 2.0 standard initiative led by the LT-Web project.[1] **Internationalization Tag Set (ITS) 2.0** recommended by the WWW Consortium provides the foundation to integrate automated processing of human language into core Web technologies. It extends the previous version of the standard with additional concepts that are designed to foster the automated creation and processing of multilingual Web content, including tags for terminology markup. TaaS terminology services enable the Web-based showcase[2] that demonstrates the application of ITS 2.0 enriched terminology annotation.

Services provided by TaaS are also registered in the **online production line PANACEA** developed by EU FP7 project that automates all steps involved in the acquisition, production, maintenance, and updating of language resources required by machine translation and other language technologies. This collaboration will facilitate the usage of the TaaS terminology services offered in other language resource processing tasks and applications.

User communities were widely involved in all phases of the platform development – from the analysis of user needs and the elaboration of usage scenarios and system interface to the evaluation and practical use of the TaaS platform. Although the platform has been in full operation only for a couple of months, it already attracted **large interest** from users.

[1] www.w3.org/International/multilingualweb/lt/
[2] taws.tilde.com/

By the time of the preparation of this report, more than 6000 users have tried the TaaS services and more than 830 users have registered to use the full features offered by TaaS. Already **more than 910 terminology projects** have been created on TaaS by its users.

The TaaS exploitation and business plan was developed to ensure the sustainability of the TaaS platform beyond the end of the European Commission co-funded project based on the freemium model to attract users and fee-based services for extended professional usage.

TaaS has made a significant contribution to the Pan-European **open language resource infrastructure** META-SHARE. A large collection of multilingual terminology is openly shared on META-SHARE consisting of more than **3 million unique term pair** candidates in EU languages that are statistically extracted from the Web (for example, multilingual websites, Wikipedia, multilingual newsfeeds etc.).

The following sections provide the description of the TaaS project results in more details.

4. Description of the main scientific and technological results

4.1. User needs analysis and functional specification

The TaaS terminology services target two main user groups: human users and machine users. Under human users, we understand language workers (for example, writers, translators, localisers, editors, content managers, domain experts, knowledge engineers, and others). Under machine users, we understand applications that process natural language content (for example, CAT tools, MT systems, search engines, and others). Within the project, TaaS addressed three major usage scenarios: usage scenario for language workers, integration with CAT to facilitate translation work, and terminology for SMT systems.

To understand terminological needs of language workers, at the beginning of the first year of the project, we conducted an online TaaS user needs survey with about 1,800 respondents. The survey included questions about professional background and terminology work (for example, importance of terminology work, time spent on it, typical problems faced, and others) as well as the needs of language workers in the area of online terminology services in order to visualise the required services and functionalities. The survey was disseminated via different channels among international associations of language workers that contribute to the exchange of information, experiences, and news in the area of terminology work (for example, IFT (International Federation of Translators) – regional centre Europe of FIT (Fédération Internationale des Traducteurs) or DTT – German Association for Terminology).

The survey confirmed that the main potential human user groups of the TaaS platform are technical writers and editors, translators, interpreters, terminologists, and domain experts. The demand on bilingual and multilingual terminology work is noticeable (see Figure 2) that confirmed the importance of providing such services as the TaaS terminology services.

Beyond the general information on the ways the potential TaaS users approach terminology work (for example, its importance, time spent, terminology collections they most frequently use, do they work alone or in cooperation with colleagues etc.) the survey results decisively influenced the elaboration of the functional specification of the TaaS platform.

For example, the survey results helped to decide on the formats, which are to be supported within the TaaS platform (see Figure 3).

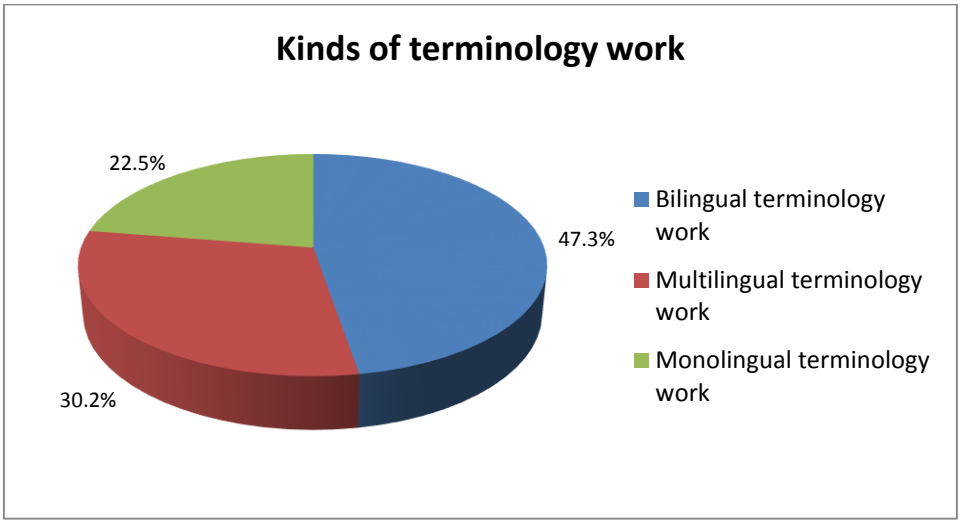


Figure 2. User needs survey: kinds of terminology work

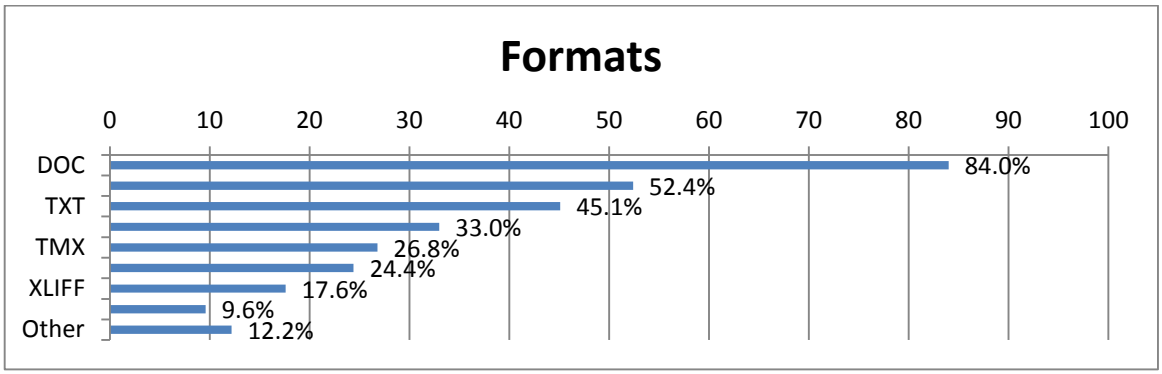


Figure 3. User needs survey: formats

As the collection of terminological data is a very comprehensive process, it was crucial to prioritise the domains that were to be initially represented in TaaS. By reference to the survey results and taking into consideration the domain coverage of existing terminology resources (we considered ETB) the prioritisation could be made[3] (see Table 1). Defining the data structure, we were eager to meet the needs and requirements of our potential users in order to make their working with TaaS as efficient and comfortable as possible. That is why we wanted to find out what terminological information they are mostly interested in (see Figure 4). That issue was also reflected in the answers of the participants as they were asked about the optimisation potential in the area of online terminology search (free-form answers). For example, it was very often mentioned that such categories like source, context, and definition are indispensable, whereby the definition should be provided for

[3] During the second year of the project, that data was further analysed and on its basis additional domains were prioritised for the open TaaS Beta.

the terms in all represented languages. The data categories implemented in the entry structure within the TaaS platform were chosen taking into consideration the survey results. In general it can be concluded that the vast majority of potential TaaS users considered terminology work to be important or very important (see Figure 5).

Table 1. User needs survey: prioritisation of domains for TaaS

Which subject fields do you cover?	Percentage of answers
Mechanical engineering !	42.9
Information technology !	33.8
Language service provider / Translation service provider / Technical documentation	26.9
Electrical engineering	24.4
Law !	16.5
Economics / Insurance business	15.1
Medicine / Dentistry !	14.8
Energy industry !	13.4
Public relations / Media	13.1
Education / Teaching&Training / University	12.6
Civil engineering	9.8
Other	9.2

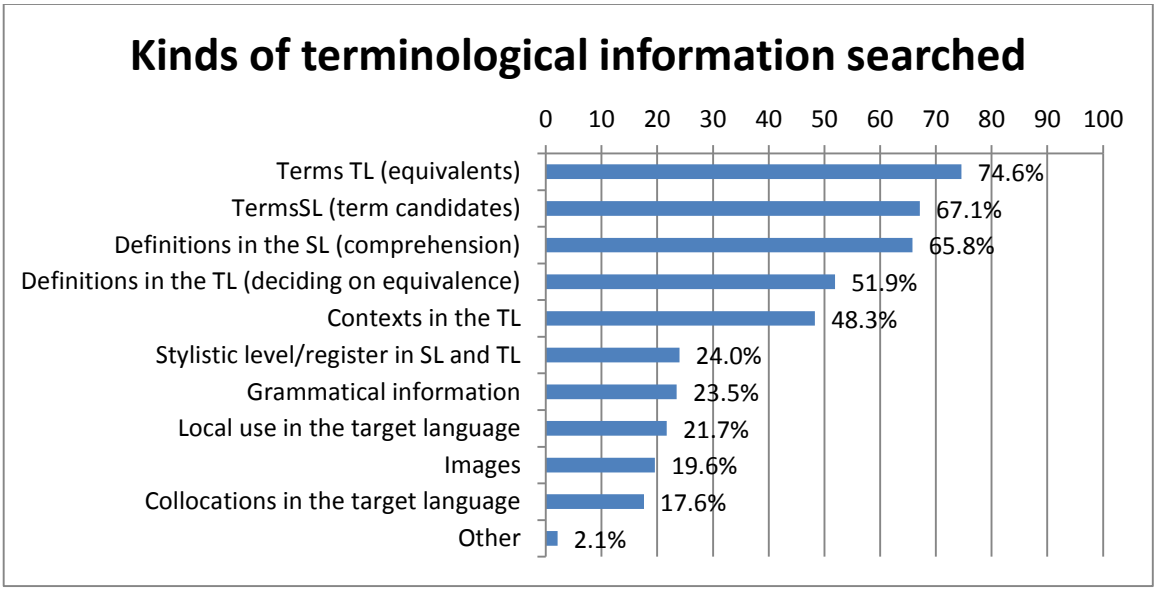


Figure 4. User needs survey: Kinds of terminological information searched for

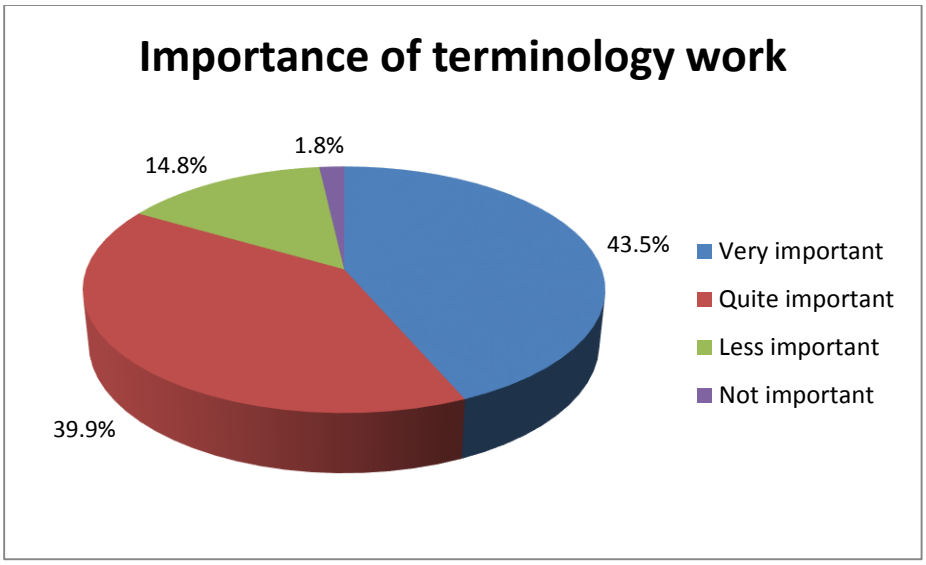


Figure 5. User needs survey: importance of terminology work

The survey participants were very much interested in online terminology services, nevertheless, they were not very satisfied with the language and domain coverage of the services they already encountered (lack of resources). Furthermore, irrespective of the working status and professional profile of the interviewees, the most frequently encountered problems mentioned while searching for terminology were poor quality of terminology found online, lack of information, and lack of reliable verification (see Figure 6).

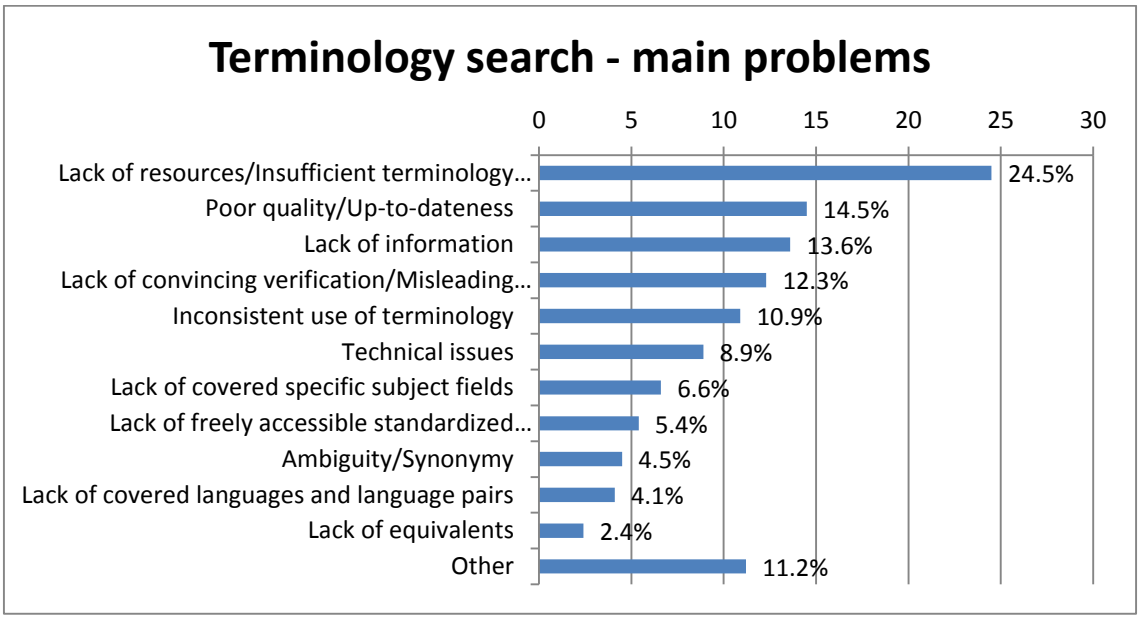


Figure 6. User needs survey: terminology search – main problems

To summarise, the areas with the greatest optimisation potential refer to the scale of terminology provided online (domains of knowledge and languages), the assurance of its

quality (verification standard), and the kinds of terminological data provided (terminological data categories). Finally, one of the main ideas of TaaS is to share the refined and approved terminology. As the survey results show, the majority of the interviewees are ready to share their terminology online (see Figure 7), although sometimes subject to certain preconditions (for example, joint contribution to the data base or strong access control).

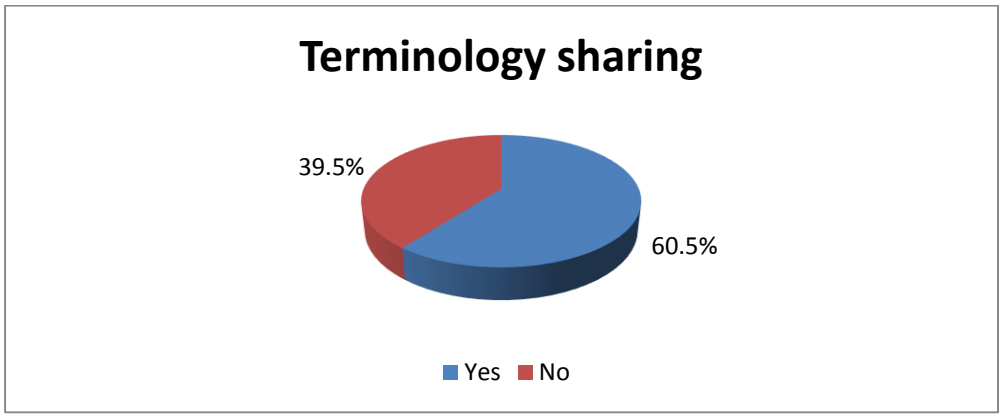


Figure 7. User needs survey: terminology sharing

4.2. Service for term candidate identification and extraction

Term candidates are identified in user-provided monolingual documents using the term tagging system Tilde Wrapper System for CollTerm (TWSC), which detects term candidates in three steps:

1. At first, documents are pre-processed using part-of-speech or morpho-syntactic taggers (and optionally lemmatisers if such exist for a language).
2. Then term candidates are extracted using linguistic filtering and statistic ranking methods. The filtering is performed with morpho-syntactic term phrase regular expressions and the ranking is performed with co-occurrence measures (for example, log likelihood, modified mutual information etc.) for terms of two or more tokens and the TF*IDF measure for unigram terms.
3. Finally, identified and extracted term candidates are marked in the user-provided documents using n-gram prioritisation and the term rankings.

During the project, TWSC was extended to the languages supported by the platform by integrating existing part-of-speech taggers (for example, the OpenNLP models for Dutch, English, French, German, Italian, and Spanish, the system by Pinnis and Goba (2011) for Estonian, Latvian, and Lithuanian, and HunPOS for Hungarian and Portuguese), building projected part-of-speech taggers for under-resourced languages using parallel corpora by Aker et al. (2014), and by generating term phrase patterns from parallel corpora following a similar approach to the part-of-speech tagger projection.

We evaluated the quality of the system for four languages in two domains (information technology and mechanical engineering). Two annotators (language specialists with focus on terminology) identified terms in two documents. The documents across all languages were on similar topics and of similar difficulty. Each of the annotators had a subjective view on what comprises a term in a context and what does not. This is because the termhood

and unithood of terms can be very ambiguous as well as subjective to the opinions of specialists who work with the terminology. Therefore, in our evaluation we used a union of their annotations and performed a precision analysis of the documents tagged by the system (see Table 2).

Table 2. Evaluation results

Language	Information Technology			Mechanical Engineering		
	Correct	Total	Precision	Correct	Total	Precision
English	213	365	58.36%	254	503	50.50%
German	198	338	58.58%	132	380	34.74%
Hungarian	147	605	24.30%	199	603	33.00%
Latvian	316	540	58.52%	331	662	50.00%

The evaluation results show that on average around 50% of the identified terms are true positives. The result may seem average. However, considering that the simultaneous identification of the term termhood and unithood is very challenging, the results are acceptable. The difficulty of the task is supported also by comparing the annotator outputs. The average agreement rate of the two Latvian annotators was only at 63.3%. In addition, the remaining term candidates are not necessarily wrong. Because of the linguistically motivated term phrase filtering, the system produces syntactically justified term candidates, which can still be useful in some application scenarios, for example, machine translation in Pinnis et al. (2012).

For users, who work on morphologically rich languages, term identification may produce very redundant term candidate lists. This can be due to the inflective nature of many languages. For example, in Czech, Latvian, Estonian etc., nouns, verbs, adjectives (and other parts of speech) may have numerous different inflected surface forms. Terms are also affected by this inflective nature and, therefore, the platform addresses this issue with term normalisation. Term normalisation is a process of transforming terms from their surface forms into their corresponding canonical forms as they are found in dictionaries and terminology resources. We use rule-based methods for term normalisation that for each of the term phrase regular expressions define a rule for term normalisation. For single-word terms, the normalised forms often correspond to the term lemmas. However, for multi-word terms, the normalised forms in many cases differ from the corresponding token lemma sequences. For example, the Latvian term “datoru tīklu” (transl. “computer network”) is normalised as “datoru tīkls”, however, the lemma sequence is different – “dators” “tīkls”. Using a rule-based approach, we can remove redundancy in the monolingual term lists.

4.3. Service for translation equivalent candidate acquisition from the Web

4.3.1. Best practice and state-of-the-art evaluation

At the beginning of the first year of the project, we reviewed best practices and state-of-the-art techniques and tools for data acquisition from the Web, monolingual term

extraction, and bilingual term alignment to identify which tools should be adopted or created for TaaS. Each tool was reviewed by considering different criteria including its language coverage, implementation or integration effort, and dependency on external tools. The analysis identified that the language coverage of existing tools were limited and that various modifications were needed to improve the language coverage to include the 25 TaaS languages. As a result, we decided on the tools to adopt or to create as shown below:

1. Corpus acquisition from the Web:
   - Parallel data acquisition: TAUS data web crawler (Ruopp and Xia, 2008)[4];
   - Comparable data acquisition: Wikipedia extractor tool (Paramita et al., 2012)[5], news retrieval tool (Aker et al., 2012)[6], and FMC Crawler (Skadina et al., 2012)[7];
2. Monolingual term extraction: TWSC (Pinnis et al., 2012)[8];
2. Bilingual term alignment: USFDTermAligner (Aker et al., 2013)[9] and MPAligner (Pinnis, 2013)[10].

These tools were chosen due to their ease of integration and flexibility to improve language coverage. Furthermore, the TaaS partners, who therefore were able to perform further modification of the tools to suit TaaS purposes, previously developed most of these tools. The adoption of these tools to further suit the project was performed in later tasks as described in the following paragraphs.

4.3.2. Parallel and comparable data acquisition for term candidate extraction

The TAUS Data Web Spider was used to acquire parallel data from the Web. As an input, the tool requires seed term lists that was compiled from the EuroVoc thesaurus. The process was performed for the following domains of knowledge: Mechanical Engineering (TaaS 1504), Information Technology (IT, TaaS 1501), Economics and Insurance Business

[4] Achim Ruopp and Fei Xia. Finding parallel texts on the web using cross-language information retrieval. In Proceedings of the second workshop on Cross Lingual Information Access (CLIA) Addressing the Information Need of Multilingual Societies, Hyderabad, India, 2008.

[5] Paramita, M., Clough, P., Aker, A. & Gaizauskas, R. 2012. Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 790-797.

[6] Aker, A. & Kanoulas, E. & Gaizauskas, R. (2012), A light way to collect comparable corpora from the web, in Proceedings of the International Conference on Language Resources and Evaluation (LREC).

[7] Skadina, I. & Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A. Babych, B., Paramita, M., Clough, P., Aker, A., Gaizauskas, R. & Glaros, N., Collecting and using comparable corpora for statistical machine translation, in Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2012.

[8] Pinnis, M., Ljubešic, N., Ştefănescu, D., Skadiņa, I., Tadić, M. and Gornostay, T. 2012. Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*. Madrid, Spain, 2012.

[9] Aker, A., Paramita, M., Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In proceedings of the Association for Computational Linguistics (ACL).

[10] Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP 2013) (pp. 562–570). Hissar, Bulgaria.

(TaaS 0300 and TaaS 0306), Medicine and Dentistry (TaaS 2200), Energy Industry (TaaS 1200), and Law (TaaS 0200) (see the TaaS domain classification). Using seed term list, we collected parallel data using the TAUS Data Web Crawler with parameter settings (Google search with inurl:language parameter and language identifier substitution matching), that in earlier experiments showed the highest yield for parallel page pairs/aligned segments per seed term. Table 3 shows parallel data extracted for several languages in the Information Technology domain (TaaS-1501).

Table 3. Data collected using TAUS Data Web Crawler

Source Language	Target Language	TaaS Domain	Seed Terms	Parallel Page Pairs	Parallel Segments
EN	DE	TaaS-1501	148	312	44969
EN	LV	TaaS-1501	148	83	18423
EN	HU	TaaS-1501	148	139	24228
EN	FR	TaaS-1501	148	199	36103
EN	PL	TaaS-1501	148	505	88997
EN	ES	TaaS-1501	148	369	66537
EN	IT	TaaS-1501	148	223	36889
EN	RU	TaaS-1501	148	85	11268
EN	PT	TaaS-1501	148	312	58467
EN	NL	TaaS-1501	148	204	34668

Comparable corpora were acquired from three different Web sources: news, generic Web sources, and Wikipedia. To extract comparable articles from news sites, we used the News Gathering tool. The tool performs a monolingual download of news articles by following the provided feed URLs. For each news article, it extracts the raw news text and saves it for the next process in the workflow – the alignment step. The alignment of gathered monolingual corpora is performed using the DictMetric tool. The tool requires RSS feeds for each language and domain. The output of DictMetric is comparable corpora. As an example, Table 4 shows the data collected using the News Gathering tool for the IT domain for several languages.

Table 4. Data collected using the News Gathering tool

Language pair	Domain	Document pairs
EN-DE	TaaS-1501	3,954
EN-LV	TaaS-1501	2,948
EN-HU	TaaS-1501	426

The FMC Crawler was used to acquire comparable articles from other generic Web sources. In order to start the corpora collection, Web domains (*not to be confused with TaaS domains of knowledge a.k.a. subject fields*) were manually collected for the TaaS languages (except Irish and Maltese, which do not have sufficient Web domains containing only Irish, nor Maltese content) so that the content found on the Web domains would belong to the particular domain of knowledge. When the Web domains were collected, we started the acquisition of comparable corpora by running the FMC Crawler for one hour for every domain of knowledge and language (i.e., FMC was executed 138 times for 1 hour) on a Web domain crawling scenario (the crawler was restricted to collecting data only from Web pages from the specified domains). The statistics of the monolingual corpora for the IT domain (TaaS-1501) for various languages is given in Table 5.

Table 5. Data collected using the FMC crawler

Lang. pair	Source				Target				DictMetric document pairs (filtered)
	Docu-ments	Unique sentences	% of mono corpus	Tokens in unique sentences	Docu-ments	Unique sentences	% of mono corpus	Tokens in unique sentences	
DE-EN	4 985	150 768	52.10%	2 187 036	1 287	60 743	60.10%	1 228 946	16 235
ES-EN	204	7 332	13.68%	98 736	342	21 356	21.13%	410 884	926
FR-EN	788	70 488	29.13%	498 183	557	25 090	24.83%	453 648	2 567
HU-EN	1 931	40 033	46.92%	496 209	1 090	51 838	51.29%	1 056 088	7 610
IT-EN	1 103	27 922	15.05%	498 401	605	33 306	32.96%	661 330	4 444
LV-EN	801	20 029	15.09%	477 597	805	41 107	40.67%	826 086	3 659
NL-EN	1 112	33 484	21.71%	438 262	758	34 768	34.40%	661 203	3 856
PL-EN	812	20 354	20.07%	291 312	717	39 428	39.01%	799 982	3 439
PT-EN	2131	56 570	81.77%	1 058 804	421	28 358	28.06%	572 198	3 997
RU-EN	2 158	63 707	23.00%	880 789	1 444	63 767	63.10%	1 282 049	7 759

The Wikipedia Extractor tool was used to acquire comparable Wikipedia documents. That was performed by downloading the Wikipedia dumps and extracting plain-text versions of the articles. Document alignment was identified using the Wikipedia interlanguage links, i.e., internal Wikipedia links that connect articles written about the same topic in different languages. The total number of articles collected for each language on Wikipedia are shown in Table 6.

Table 6. Data collected from Wikipedia extractor tool

Language Pair	Document Pairs	Language Pair	Document Pairs
BG-EN	106,708	IT-EN	703,915
CS-EN	180,271	LT-EN	86,282
DA-EN	121,529	LV-EN	38,263
DE-EN	812,640	MT-EN	2,696
EL-EN	62,161	NL-EN	622,564
ES-EN	645,305	PL-EN	627,642
ET-EN	70,424	PT-EN	533,158
FI-EN	232,134	RO-EN	175,139
FR-EN	888,485	RU-EN	548,825
GA-EN	21,307	SK-EN	135,102
HR-EN	87,186	SL-EN	82,599
HU-EN	163,806	SV-EN	445,854

4.3.3. Service for term candidate extraction

TWSC is a console application that allows tagging terms in plaintext documents using state-of-the-art linguistically and statistically motivated term extraction methods. TWSC was initially developed in the ACCURAT project for the Latvian and Lithuanian languages. The original documentation of TWSC included in the documentation of the ACCURAT Toolkit (section 4.1) contains a detailed description of the term extraction and tagging methods applied in the system. During the ACCURAT project, it was also adapted for term tagging in English and German language documents. In the TaaS project, the TWSC tool was further adapted for the languages included in the scope of the project. TWSC is used for monolingual term extraction on parallel and comparable data prior to the bilingual term alignment process.

For each language, TWSC requires three linguistic resources in order to achieve higher term tagging performance in terms of recall and precision. The first is a reference corpus statistics in the form of an Inverse Document Frequency (IDF) list. The IDF lists for the project's languages within the TaaS project were created using the Wikipedia corpora. It also requires a stop word list, which was created by extracting the 200 lowest ranked words from the IDF list, i.e., top 200 most common words. Finally, TWSC requires a part-of-speech (POS) tagger and a corresponding valid term phrase pattern list (term grammar). For languages that do not have available POS taggers, we used a projection technique to create the POS tagger and the term grammar Aker et al. (2014), enabling TWSC to identify terms for all 25 TaaS languages. Figure 8 shows an example of the TWSC output.

```
From a <TENAME SCORE="0.32" MSD="JJ NN" LEMMA="biological perspective">biological
perspective</TENAME>, the evidence is strongly persuasive that physical <TENAME
SCORE="0.08" MSD="NN" LEMMA="activity">activity</TENAME> reduces the occurrence of these
leading <TENAME SCORE="0.04" MSD="NNS IN NN" LEMMA="cause of death">causes of
death</TENAME> (discussed in the individual chapters on these diseases); thus, it is also
biologically plausible for physical <TENAME SCORE="0.08" MSD="NN"
LEMMA="activity">activity</TENAME> to postpone the occurrence of <TENAME SCORE="0.64"
MSD="JJ NN" LEMMA="all-cause mortality">all-cause mortality</TENAME>. (Because we all die
eventually, when the phrase "lower risk of <TENAME SCORE="0.64" MSD="JJ NN" LEMMA="all-
cause mortality">all-cause mortality</TENAME>" is used in this chapter, it refers to lower
risk during the <TENAME SCORE="0.15" MSD="NN IN NN" LEMMA="period of follow-up">period of
follow-up</TENAME> in a study; i.e., postponed <TENAME SCORE="0.15" MSD="NN"
LEMMA="mortality">mortality</TENAME>.)
```

Figure 8. Example of a term-tagged document by TWSC

4.3.4. Service for bilingual term alignment

After all terms are identified in the document, bilingual term alignment is performed using MPAligner (Pinnis, 2013)[11], which is a context independent term mapping tool developed by Mārcis Pinnis (a public version of the mapper in available on GitHub). The tool maps two terms (in two different languages) using a method that tries to find the maximum content overlap between the two given terms with the help of maximised character alignment maps. The approach (character level content overlap analysis) allows the mapper to map multi-word terms and terms with different numbers of tokens in the source and target language parts. The mapper was specifically designed to address term mapping between European languages (including languages with different alphabets based on Latin, Cyrillic, and Greek) and it allows the integration of linguistic resources to increase recall (while maintaining the same level of precision) of the mapped terms. This tool supports the term mapping between the TaaS 25 languages. The output of the tool is shown in Figure 9.

```
lv      mobilajām ierīcēm    en      mobile devices       TaaS-1501
        http://taas.eurotermbank.com/TILDE-MPAligner/USFD-News-Week3-IT-Corpus/v3
        0.992336      A-fpdp-y------------------l- N-fpd---------n-----------l-
        mobils ierīce mobilā ierīce A-fsnp-y------------------l- N-fsn---------n--
---------l-   starp Samsung SMART televizoriem un mobilajām ierīcēm. Turklāt TV
Discovery spēj       JJ NNS mobile device mobile device JJ NN  tablets , and use
their mobile devices as remotes to control their
        /home/marcis/TILDE/TAAS/USFD_RSS_NEWS_WEEK3/it/en-lv/lv/lv2826.txt
        /home/marcis/TILDE/TAAS/USFD_RSS_NEWS_WEEK3/it/en-lv/en/enFile265.txt
```

Figure 9. Output of MPAligner for Latvian-English terms

4.3.5. Bilingual Term Extraction System (BiTES)

Existing tools were adopted to address the TaaS requirements. New tools for the processing of the input text, such as new POS tagger wrappers, document classifier, term normalisers, and term alignment tools, were also developed in the TaaS project. All tools were integrated in four workflows (which differ in the way the data acquisition process is performed) within BiTES. Assuming that the required input data of each data acquisition process is provided,

[11] Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)* (pp. 562–570). Hissar, Bulgaria.

these workflows can be run automatically either by a system administrator or iteratively in order to gather more data. The four workflows are shown in the following figures: the parallel Web data acquisition workflow (see Figure 10), the news data acquisition workflow (see Figure 11), the FMC data acquisition (see Figure 12), and the Wikipedia data acquisition (see Figure 13). Figure 14 visualises the BiTES workflows for multilingual terminology extraction from comparable corpora acquired from the Web.

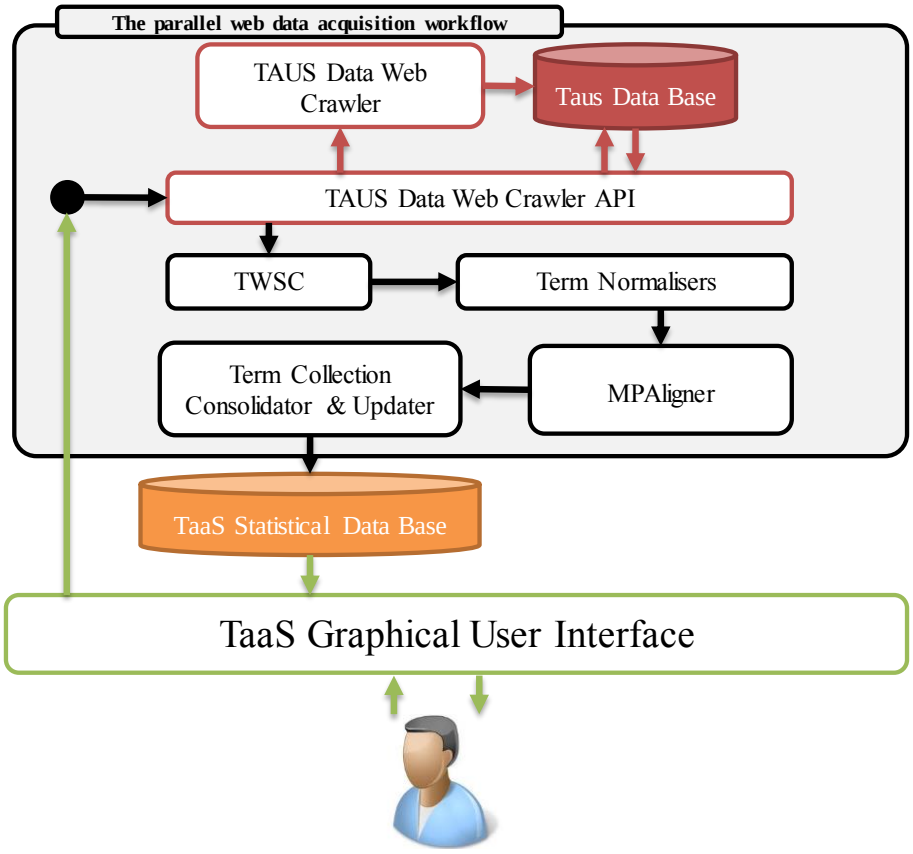


Figure 10. The parallel Web data acquisition workflow

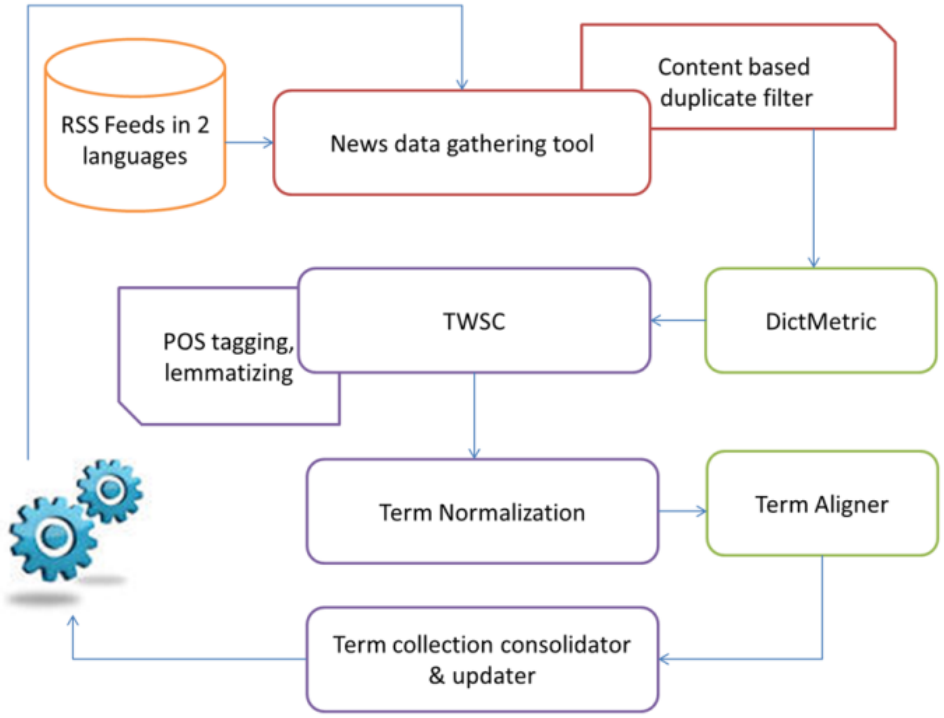


Figure 11. The news data acquisition workflow

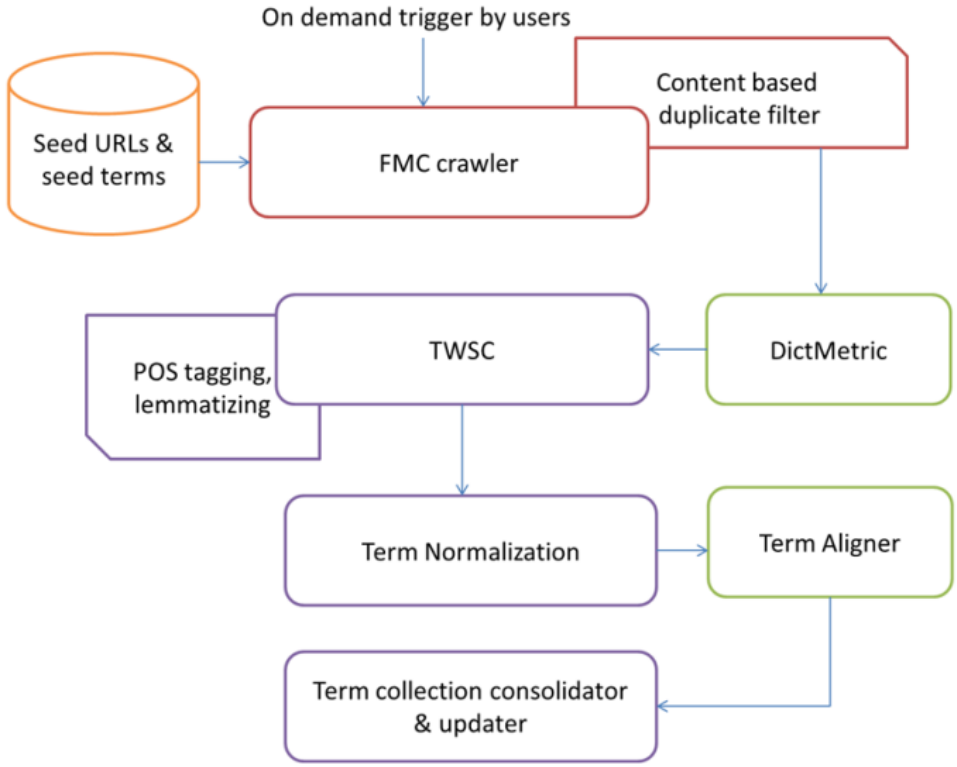


Figure 12. The FMC data acquisition workflow

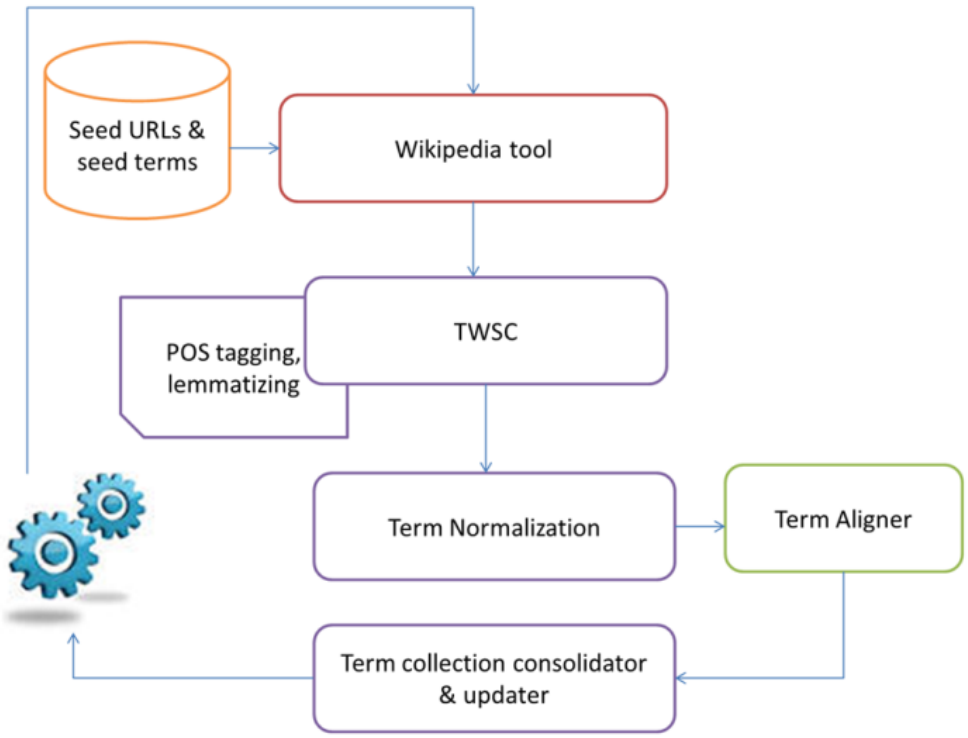


Figure 13. The Wikipedia data acquisition workflow

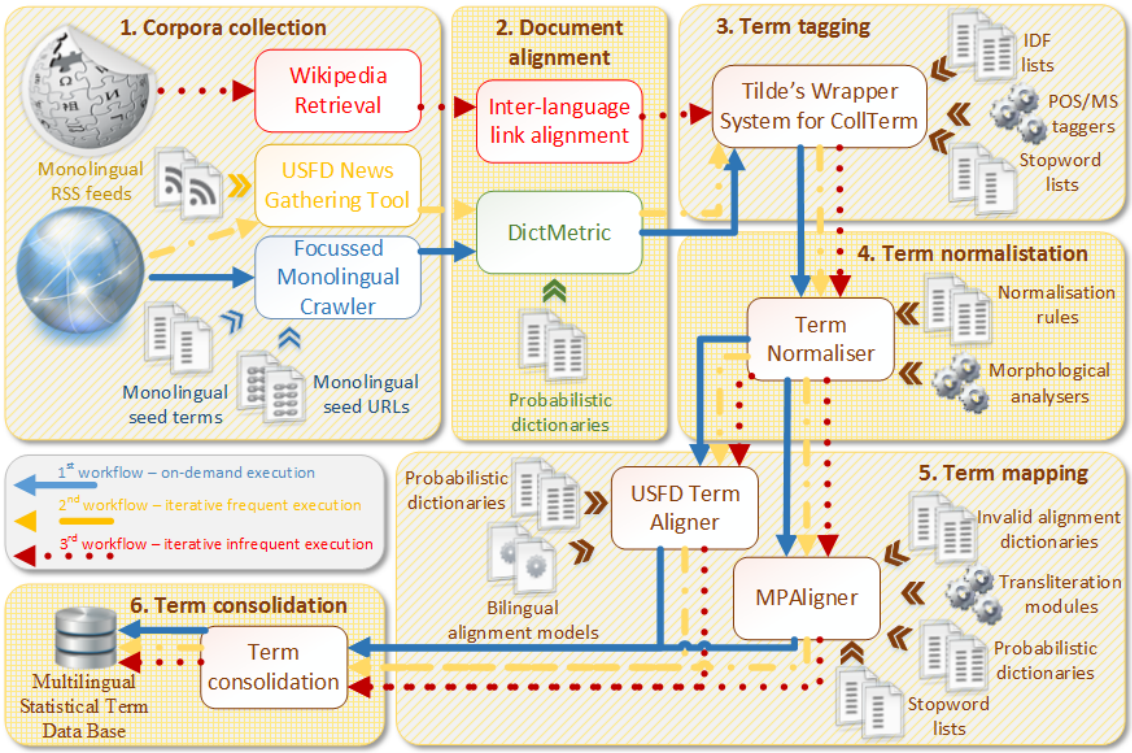


Figure 14. Workflows for multilingual terminology extraction from comparable Web corpora

4.3.6. Evaluation of term extraction and alignment

Automatic evaluation of different TaaS components was performed by evaluating recall of term extraction and term alignment of EuroVoc terms using TWSC and MPAligner. Manual evaluation was performed by evaluating a set of randomly sampled data, i.e., pre-classified documents, candidate terms, and candidate term pairs. In addition, the quality (accuracy) of term pairs between different workflows was evaluated. As a result, the BiTES system achieved between 87% accuracy (FMC) to 98% accuracy (parallel data) (see Table 7). We also identified the ratio of term pairs per document pair and found that both parallel data and Wikipedia are promising sources for bilingual term extraction (see Table 8).

Table 7. Accuracy of term pairs between different workflows

Languages	Parallel Data	News	FMC	Wikipedia
All languages	98%	90%	87%	94%
CS-EN	-	93%	91%	90%
DE-EN	97%	95%	97%	90%
ES-EN	98%	94%	90%	97%
LT-EN	98%	83%	82%	98%
LV-EN	98%	84%	85%	95%

Table 8. Ratio of term pairs per document pair

Workflows	Ratio of terms
Wikipedia workflow	5.1 term pairs per document pair
FMC workflow	0.22 term pairs per document pair
News workflow	3.5 term pairs per document pair
Parallel data workflow	0.5 term pairs per parallel text segments

4.3.7. TaaS Statistical Database

Multilingual terminology automatically extracted using BiTES from parallel and comparable data acquired from the Web is stored in the TaaS Statistical Database (SDB). SDB operates as a backup resource for bilingual terminology lookup if authoritative resources (i.e., term banks like ETB, IATE as well as user-created terminology collections accessible on the TaaS platform) do not return term translation equivalents during the bilingual terminology extraction process. SDB allows storing terms and cross-lingual links between terms in different languages, thus creating a multilingual resource (see Figure 15).

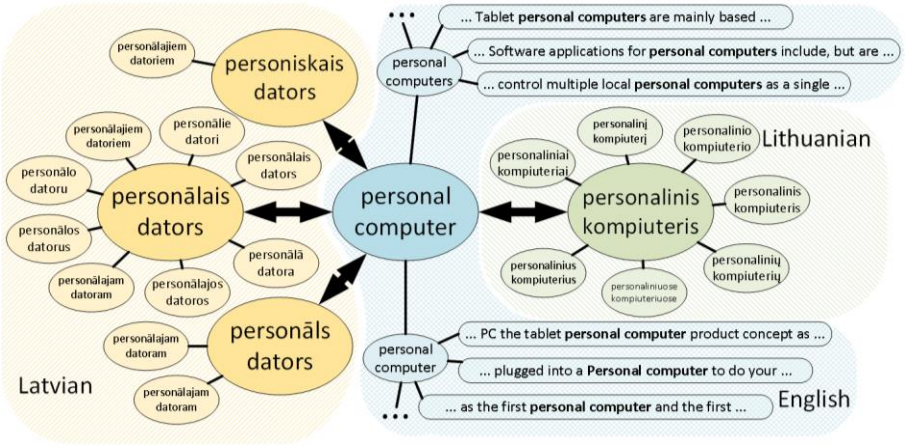


Figure 15. Visualised example of terminological data from SDB

For each term, depending on the linguistic tool support for each language, SDB can provide its canonical form, different surface forms (which are important for morphologically rich languages), morpho-syntactic information, lemma sequence, information about sources (for example, the BiTES workflow that produced the term pair, the tools involved in the extraction process, the documents, from which the term was extracted), and contexts, in which the term was identified. Terms in SDB are grouped by canonical forms in order to provide a list of unique translation candidates for user extraction projects.

SDB is a constantly growing resource as the BiTES workflows are iteratively providing new bilingual term collections in the form of TSV (Tab-Separated Values) documents. SDB is then iteratively updated with the new terminology collections. The statistics of SDB (at the time of writing this report) in terms of unique term surface forms per language is given in Table 9 at the end of this section. BiTES provides four main workflows for automatic bilingual terminology acquisition from the Web. During the TaaS project, we actively applied the workflows in order to acquire initial terminological data from the Web for the TaaS platform so that users can instantly benefit from the TaaS platform services. The statistics of the bilingual terminology extracted from the Web for SDB using the four different BiTES workflows is given in Figure 16.

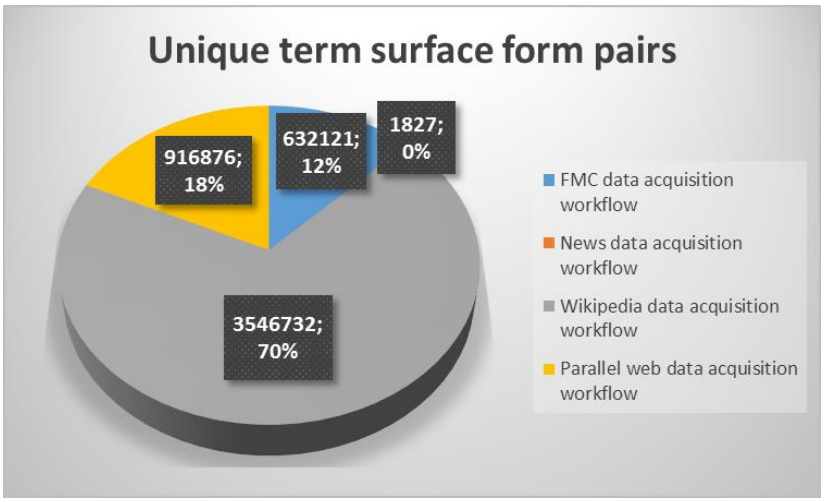


Figure 16. Statistics of the BiTES workflow data for SDB

Using the BiTES workflows, we can acquire bilingual term pairs processing a single language pair at a time. Bilingual term pairs, when integrated in SDB, are transformed into a multilingual statistical term database (for which the statistics is given in Table 9 below). However, the statistics of the automatically extracted bilingual terminology for SDB using all four TaaS workflows with respect to different language pairs is given in Figure above. In total, during the TaaS project, we acquired over 5 million unique term surface form pairs using the four different BiTES workflows (see Figure 17).

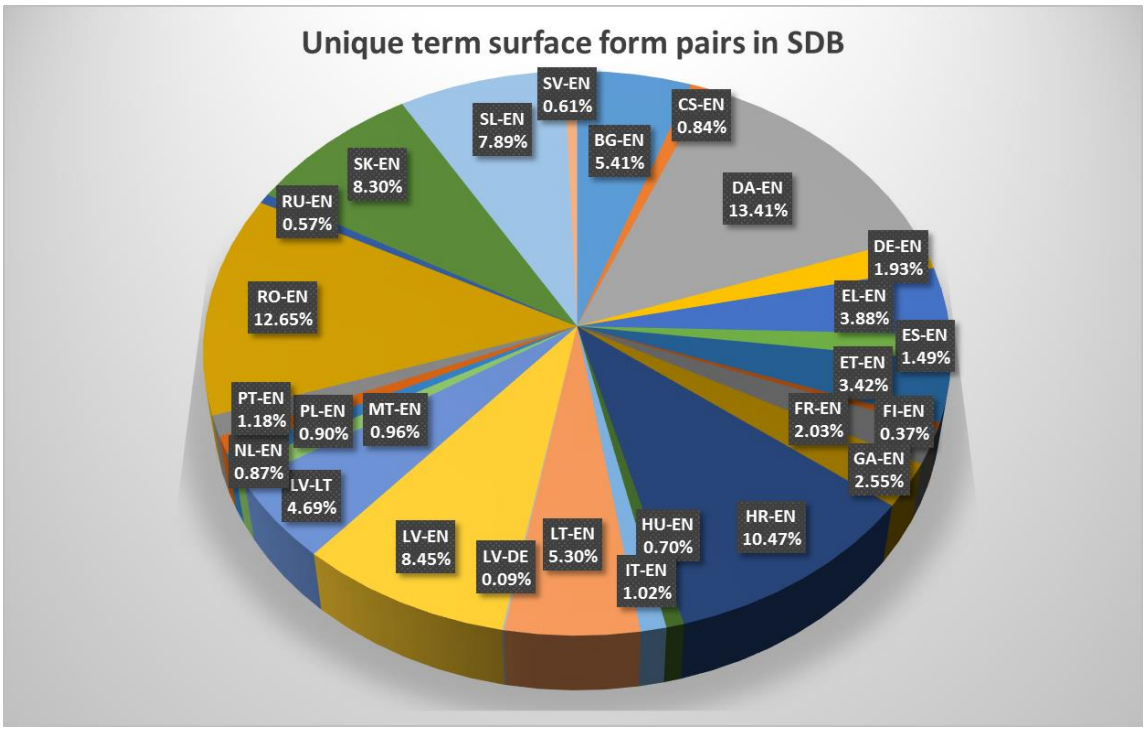


Figure 17. Statistics of unique term surface form pairs in SDB for different language pairs

Table 9. SDB Statistics – unique surface forms of terms in different languages and domains

Language	Agriculture and foodstuff	Arts	Economics	Energy	Environment	Industries and technology	Law	Medicine and pharmacy	Natural sciences	Politics and Administration	Social sciences	Grand Total
Swedish			6 296	5 768		5 849	5 848	2 325				**26 086**
Bulgarian	5 941	8 851	15 696	10 302	3 878	26 495	14 708	5 628	2 679	36 837	57 279	**188 294**
Czech			8 447	8 079		6 634	8 154	3 635				**34 949**
Danish	14 357	36 615	32 500	15 992	6 823	76 768	36 642	5 549	4 152	97 645	219 148	**546 191**
German			10 333	17 477		23 230	9 861	17 832				**78 733**
Greek	2 913	5 464	13 891	9 095	2 248	22 567	13 212	5 715	1 600	19 658	32 857	**129 220**
English	56 874	130 793	136 586	101 911	47 762	311 114	172 740	66 001	27 593	330 024	679 620	**2 061 018**
Spanish			7 941	13 960		12 808	10 653	12 996				**58 358**
Estonian	3 548	6 707	8 061	4 834	2 323	15 832	8 686	2 783	1 640	22 127	40 555	**117 096**
Finnish			3 813	2 375		4 213	4 441	1 834				**16 676**
French			12 547	17 009		14 669	13 599	16 450				**74 274**
Irish	3 526	6 238	2 651	994	946	7 586	25 731	719	853	18 596	49 902	**117 742**
Croatian	8 979	30 026	17 103	13 877	6 919	50 495	28 189	9 216	3 435	67 340	186 455	**422 034**
Hungarian			5 907	8 230		5 990	5 653	3 435				**29 215**
Italian			7 339	8 579		9 670	7 711	6 331				**39 630**
Lithuanian	5 466	6 686	13 278	18 699	18 120	27 145	12 187	5 277	1 381	38 719	44 100	**191 058**
Latvian	3 161	3 738	10 214	27 549	30 674	48 338	9 871	19 595	1 092	36 825	35 554	**226 611**
Maltese	425	1 372	4 950	6 338	308	4 577	5 431	2 202	273	4 184	6 706	**36 766**
Dutch			8 306	7 686		8 853	7 017	3 616				**35 478**
Polish			6 633	9 272		7 889	6 829	6 548				**37 171**
Portuguese			9 998	8 209		13 697	9 390	5 670				**46 964**
Romanian	10 384	26 401	27 867	15 122	7 892	62 158	28 333	6 336	7 443	96 748	188 117	**476 801**
Russian			3 201	4 526		5 864	5 410	3 933				**22 934**
Slovak	6 531	18 101	22 828	14 090	5 822	53 273	21 299	5 607	6 107	41 915	131 710	**327 283**
Slovenian	9 319	16 499	24 686	15 699	5 866	54 296	21 083	8 400	5 260	49 114	107 532	**317 754**
Grand Total	**131 424**	**297 491**	**421 072**	**365 672**	**139 581**	**880 010**	**492 678**	**227 633**	**63 508**	**859 732**	**1 779 535**	**5 658 336**

4.4. Service for terminology sharing and application

Terminology resources are important not only for language workers but also for various language processing applications ("machine users") such as CAT tools and MT systems. For the usage of terminology services and terminological data by external engines, the TaaS platform provides the TaaS API. The TaaS API allows for authorised and anonymous access to the TaaS terminology services.

To integrate the TaaS terminology services with a CAT tool or other external application, a use creates an API key. The API key is created within the CAT tool, and the user will be redirected to the TaaS API key section that provides an interface for managing the user's API keys. When the user has signed up on TaaS and his/her API key is retrieved and stored by an external application, the user is able to access TaaS collections that he/she has permission to view and to work with. If the user has not signed up on TaaS, he/she is able to access the TaaS public collections. The API-level integration is currently implemented by memoQ and OmegaT CAT tools.

For MT, before using the TaaS API, an MT system developer has to submit a request for a machine user (a key and a password, with which the system can authenticate itself within the TaaS platform). Then, similarly to a CAT tool, each user of an MT system has to be signed up on the TaaS platform and has to request a user key that the machine user (the MT system) can then use to impersonate as a human user (provided that the user grants rights to the MT system) and access his/her private term collections. The API-level integration is currently implemented by the LetsMT SMT system. TaaS is used to create project specific terminology resources for dynamic adaptation of SMT systems.

Refined, enriched, and approved terminology can be exported in one of the TaaS export formats (see section 3) and used in other working environments. Approved terminology can also be used in other terminology projects by the user(s), who owns the data, as well as by other users, provided that the terminology collection is public.

4.5. TaaS platform

During the first year of the project, the TaaS Shared Term Repository (STR) was designed, developed, and populated with initial terminological data in four domains and for eight languages. Initial terminological data was extracted from parallel and comparable corpora acquired from the Web. The data model for STR was proposed on basis of the status of relevant international standards for data exchange and data modelling. STR provides an online access point for the interface layer (human and machine users of the TaaS terminology services), the application logic layer (the TaaS terminology services, user management, and others), and the data storage layer (the storage of data at different levels, for example, user data, project data, and terminology).

During the second year, the TaaS infrastructure of cloud terminology services was established integrating all the components into the integrated TaaS system. The integrated TaaS system included all the platform facilities (for example, facilities for terminology clean-up and sharing, advanced deployment of the TaaS cloud infrastructure for optimal performance of terminology services and their robustness, the TaaS system automatic scaling with respect to its workload, iterative updates of the TaaS Web-based GUI with

monthly releases starting with the first public release of the TaaS open Beta in November 2013, the mechanism for terminological data normalisation and consolidation before its import into the TaaS Shared Term Repository, and other facilities).

4.6. TaaS portal

TaaS is a cloud-based platform providing language workers and other interested groups with facilities for the acquiring, processing, sharing, and reusing of multilingual terminological data as well as for searching terminology. TaaS can be accessed at demo.taas-project.eu or termunity.com. Without being logged in, a user can search terminology and process a text for terminology identification and lookup. The homepage contains the general information about the platform starting with the provided terminology services (see Figure 18).

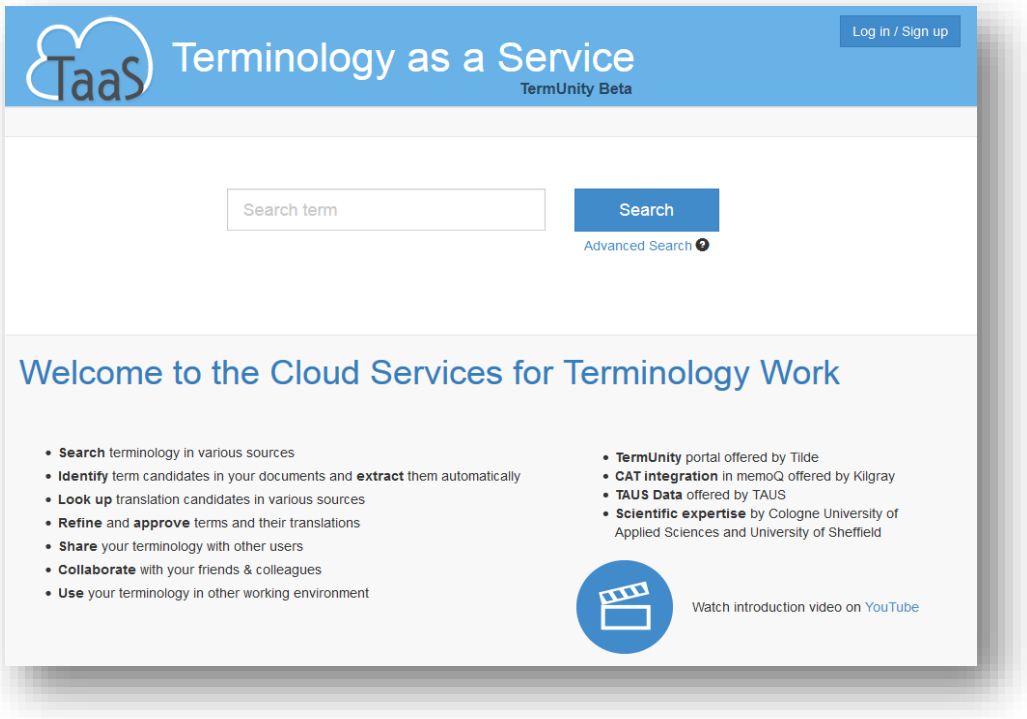


Figure 18. TaaS homepage: terminology services

4.6.1. Facilities for terminology acquisition

The registered user can benefit from the whole range of the terminology services offered by TaaS and available via the TaaS Web-based GUI. These services can be subdivided into the following groups:

- Administrative unit dealing with the services around the user account;
- Services dealing with terminology data management;
- Services dealing with the integration with CAT tools and MT systems;

- Linking and feedback facilities.

When logged in into the platform, the user can view and manage the user account: view his/her account details, edit the personal information, or delete the account. Under "Edit". The user can modify the information on his/her real name, profession, country, and working languages.

The terminological data is processed within projects. The user can access existing projects, where he/she has an access rights or create a new project. The following functionalities are provided for the user's work with projects (see Figure 19 and Figure 20):

- Documents: the user can upload documents for processing in various formats.
- Extraction: the user can select documents for extraction, define tools for extraction and sources for target translation lookup.
- Terms: the user can view and edit the extraction and the translation lookup results.
- Visualisation: the user can view the extraction results in the processed text.
- Sharing: the user can share his/her projects.

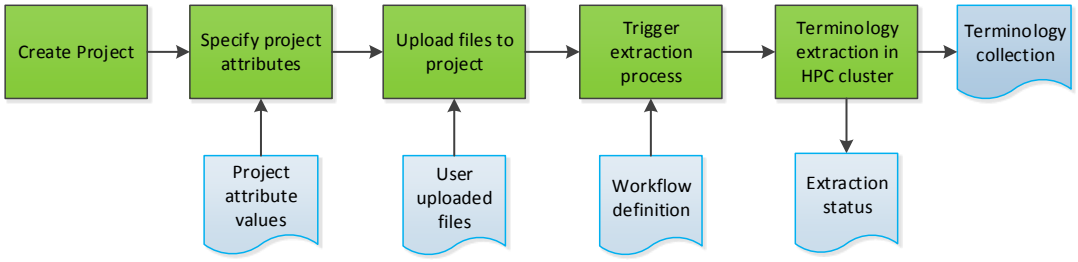


Figure 19. TaaS project workflow

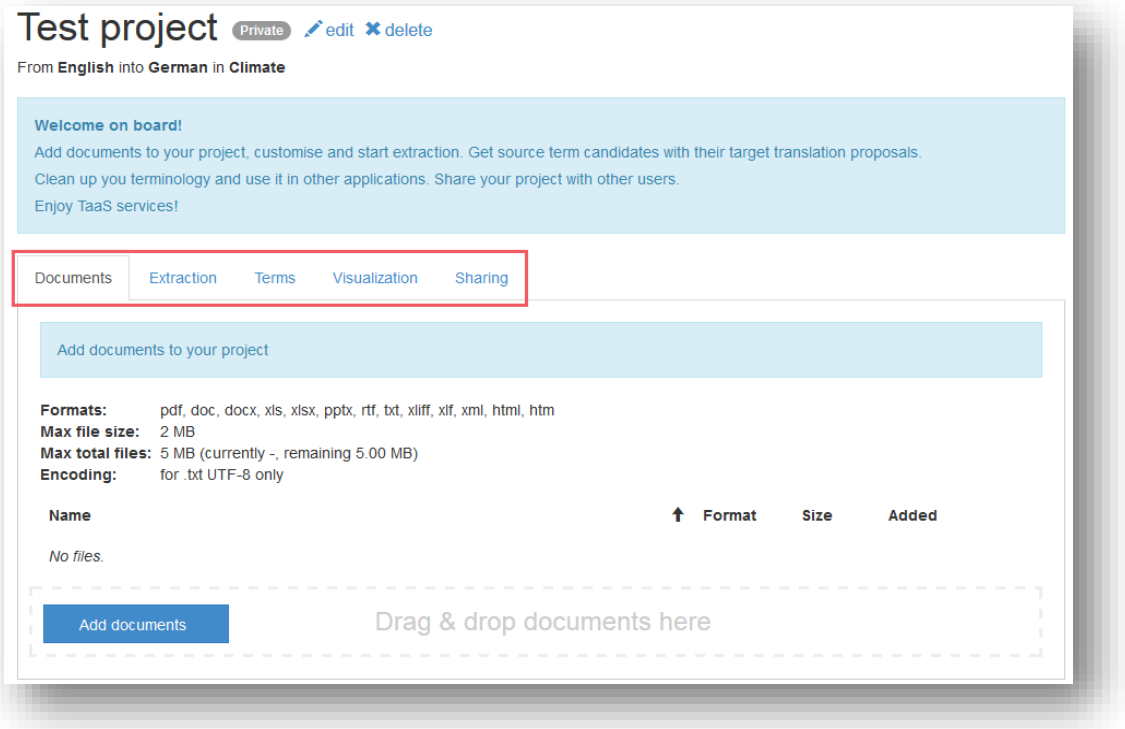


Figure 20. TaaS "Working with a project"

For term extraction, the user can specify the sources for the translation candidate search:

- External term bases and term banks (ETB, IATE – these sources can be considered as more reliable).
- External sources that store crowd-sourced data (TAUS).
- Terminology refined, enriched, approved, and then published by the TaaS users, it is stored in the TaaS STR.
- Terminology automatically extracted from the Web and stored in the TaaS SDB (this source might be considered as more noisy).

4.6.2. Facilities for terminology refinement, enrichment, and approval

Having processed a text, a list of term candidates and translation equivalent candidates is generated. By default, the user can see term candidates sorted by their translation equivalent candidates. It is possible to sort/filter terminology according to the first letter ([a]), page-wise ([b]), hide term candidates according to the source the translation comes from ([c]), hide candidates with no approval ([d]) or with no translation ([d]) (see Figure 21).

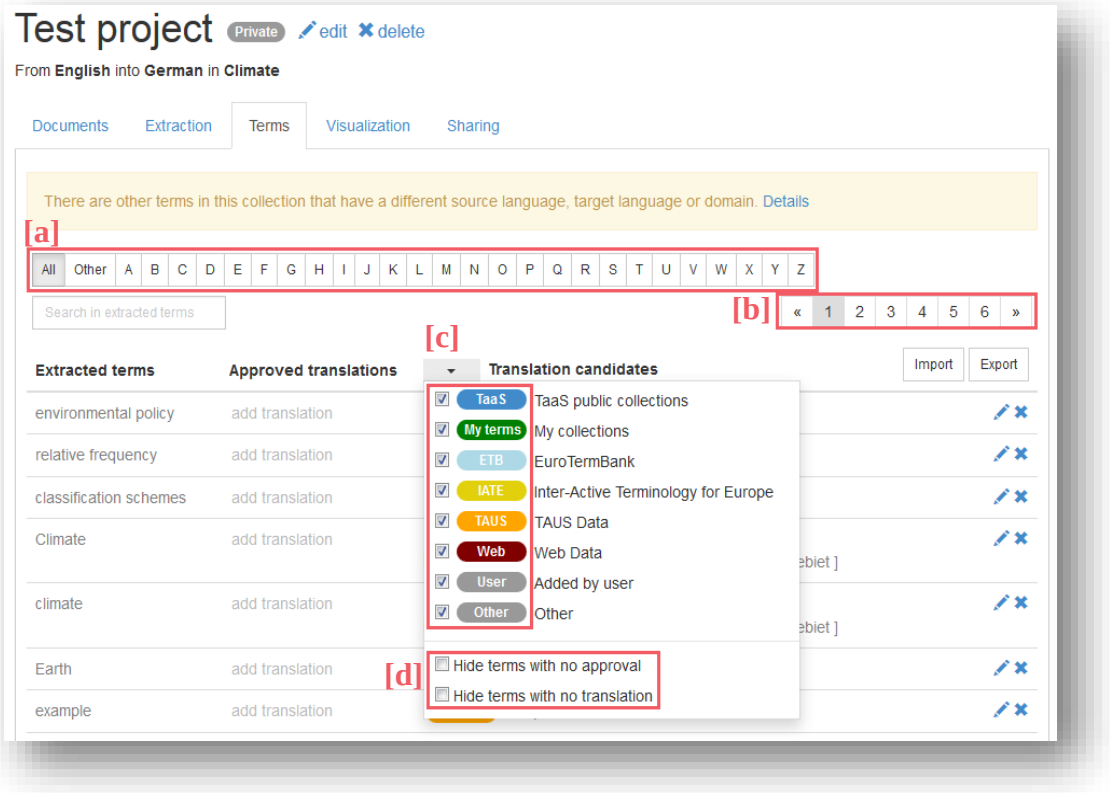


Figure 21. TaaS "Sort term candidates"

In the latest version of the TaaS platform, the list of term candidates can be also sorted alphabetically and by frequency a term is used in the user's document(s).

The approval process is very simple: with a single click, the user approves a term candidate and/or its translation. The Term Entry Editor is available for the refinement and enrichment process: to edit a term, add a definition/note, and complement some other data categories, for example, term type, register, or geographical usage.

In addition, the user can import and export his/her terminology in one of the TaaS formats supported by TaaS.

The list of term candidates can also be searched and new source terms can be added.

Although a bit of context is provided during the extraction process and can be viewed when putting the cursor on a term candidate or when editing the entry, it also might be useful to see the extracted candidate in the context of the whole text. For this purposes, the visualisation option of the processed document is provided (Figure 22 and Figure 23). Identified terms are highlighted, and the user can get detailed information and all the contexts by putting the cursor on a term candidate.

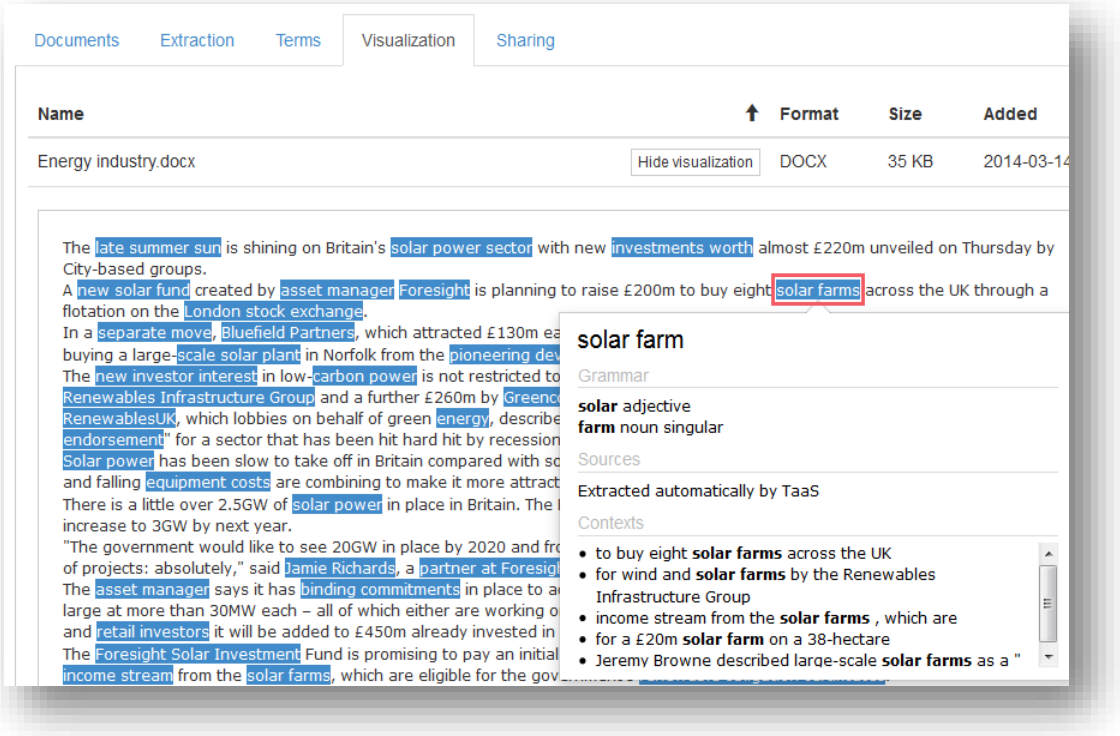


Figure 22. TaaS visualisation of extraction results (1)

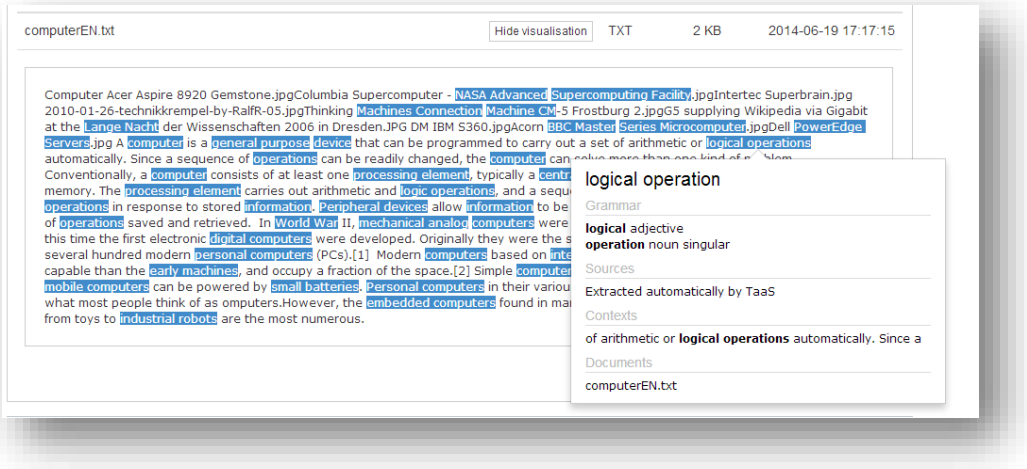
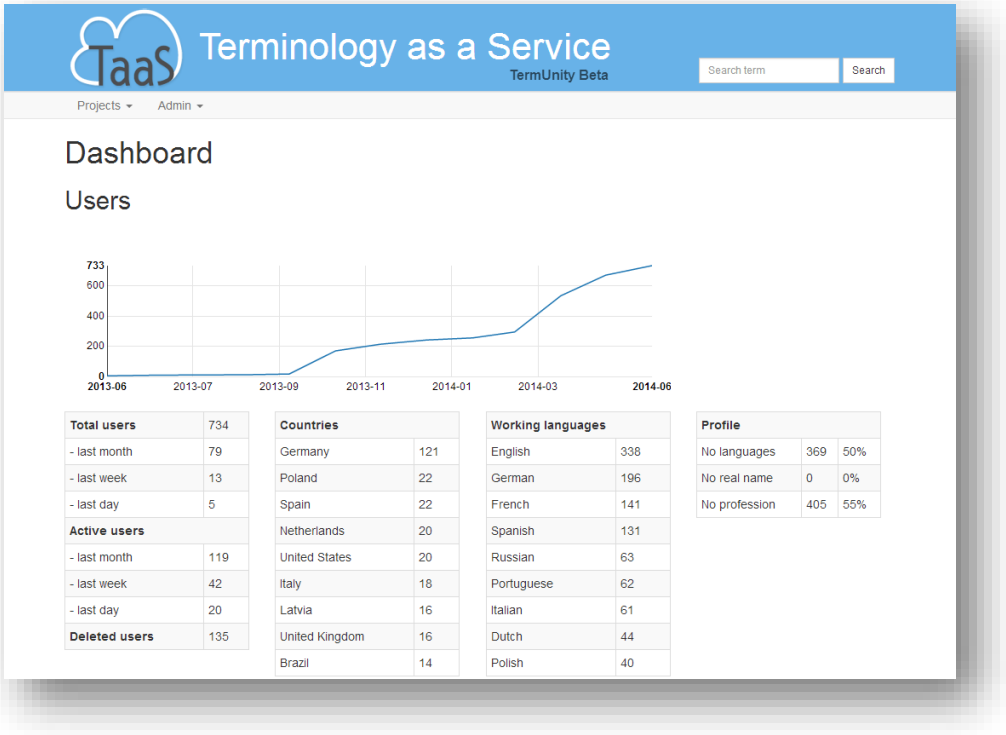


Figure 23. TaaS visualisation of extraction results (2)

The user can share his/her projects with other users assigning different roles to them. For this purpose, the user has to enter the e-mail address of the user in question and to select the role (Reader, Editor or Administrator).

To analyse the activity of the TaaS users and the usage patterns of the TaaS services, the administration dashboard is available for the platform administrators that shows the reports of usage statistics from different perspectives (Figure 24).



Total users	734
- last month	79
- last week	13
- last day	5
Active users	
- last month	119
- last week	42
- last day	20
Deleted users	135

Countries	
Germany	121
Poland	22
Spain	22
Netherlands	20
United States	20
Italy	18
Latvia	16
United Kingdom	16
Brazil	14

Working languages	
English	338
German	196
French	141
Spanish	131
Russian	63
Portuguese	62
Italian	61
Dutch	44
Polish	40

Profile		
No languages	369	50%
No real name	0	0%
No profession	405	55%

Figure 24. TaaS Dashboard for usage analytics

4.6.3. CAT tool memoQ usage features

The TaaS services are easily accessible from the CAT tool memoQ working environment (see Figure 25 and Figure 26). The integration in memoQ is implemented via the TaaS API using the following methods:

- Authentication and Authorisation;
- Term lookup;
- Retrieve entry;
- Add new entry / Edit entry;
- Retrieve list of supported languages and domains.

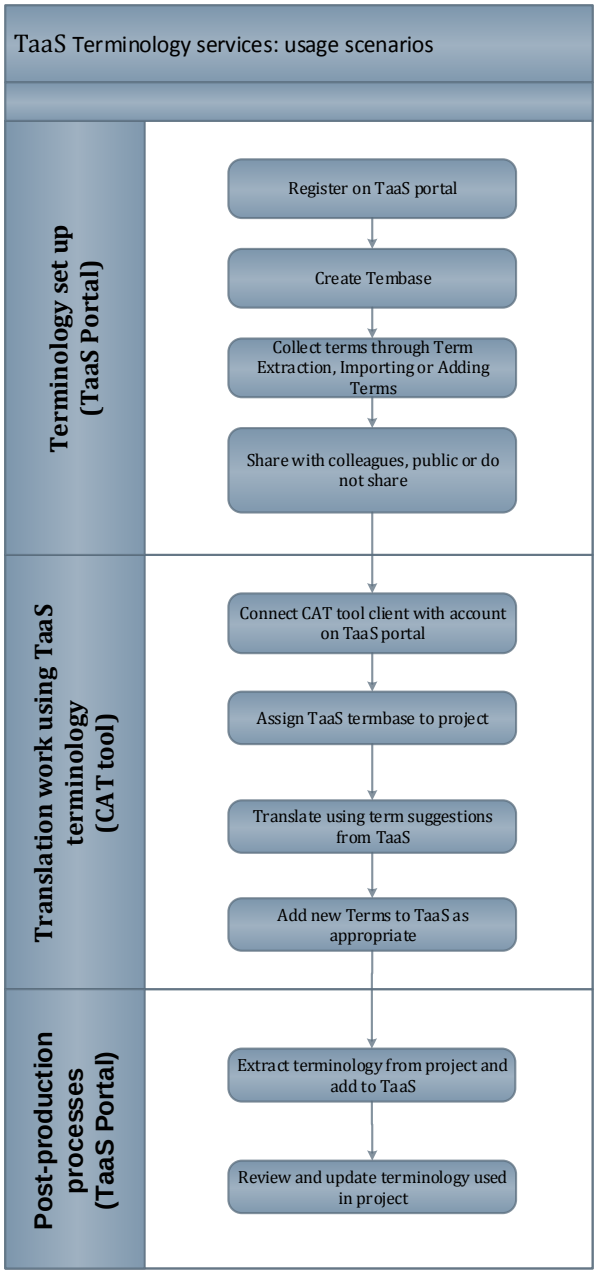


Figure 25. Usage Scenarios for CAT tool users

4.7. Improving statistical MT with terminology data

TaaS made significant efforts in researching terminology integration in statistical machine translation (SMT) systems. When performing automated translation with SMT systems it is important that the systems are able to provide support for correct handling of terminology by assuring that the right terms are used in the translation and that their usage is consistent across the translated text.

For machine translation systems, the first requirement is difficult to achieve, because the context (or more precisely, the lack of enough context) may not always allow identifying

the correct translations of terms. The second requirement challenges SMT systems more than rule-based MT systems as the statistics of large amounts of data is difficult to control if not constrained by means of, e.g., bilingual term glossaries or translation model or language model domain adaptation techniques. If SMT systems are not developed and "taught" to understand terminology, ambiguous or unknown contexts in the parallel training data may result in the selection of incorrect translation hypotheses because of higher contextual likelihood. Therefore, in the TaaS project we have spent large efforts in order to research and develop methods for integration of user tailored terminology glossaries into SMT systems to achieve domain adaptation and produce better quality translations. Figure 26 shows the overall conceptual design of the terminology integration scenarios investigated in the TaaS project.

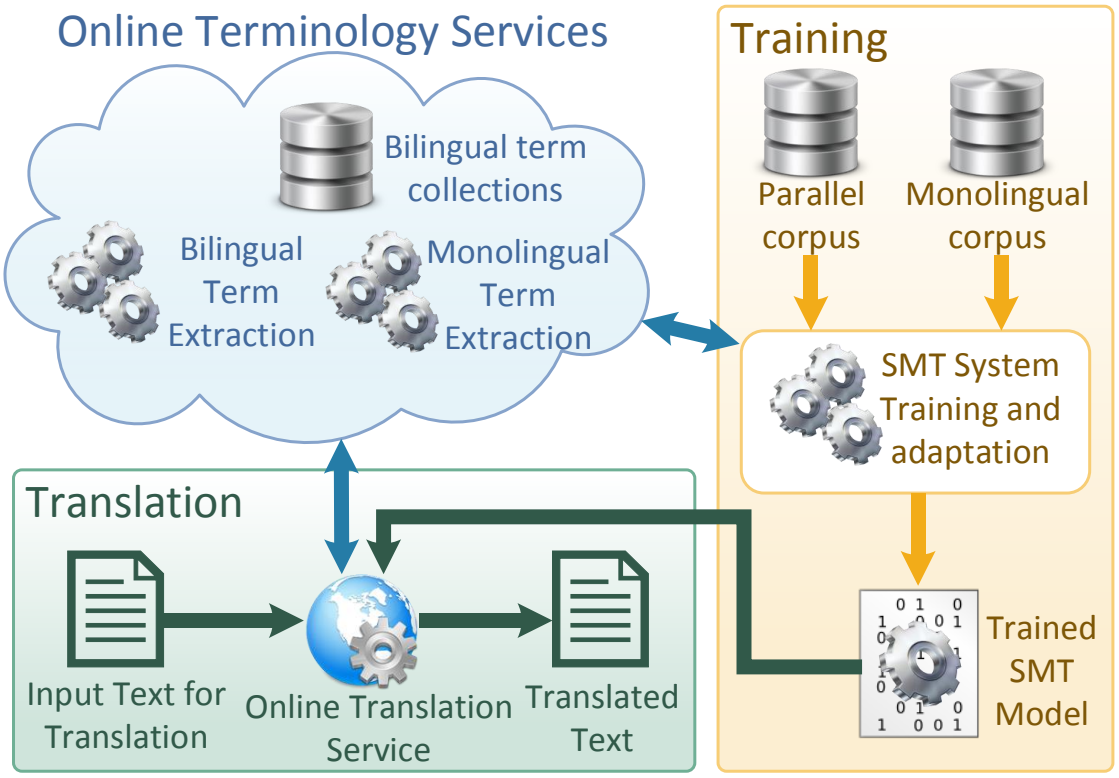


Figure 26. The conceptual design of terminology integration in SMT systems

The publicly available TaaS API has been designed so that it provides the necessary functionality for SMT systems for bilingual term collection acquisition and terminology identification in either SMT system training data or the translatable content that is passed to the SMT systems for translation.

For SMT system developers, we have developed novel approaches for bilingual terminology integration in SMT system training. The methods have been designed for the LetsMT platform, however, they can be reused by other SMT platforms that are based on the Moses SMT toolkit with minimal integration efforts. The methods allow:

- Performing SMT system translation model adaptation through:

    - Reduction out-of-vocabulary rate of term translations and improvement of in-domain term word alignment quality by adding the bilingual terminology to the parallel data. Automatic evaluation results show that we can improve SMT quality in terms of BLEU by up to 7.5% over baseline systems by enriching SMT training data with in-domain terminology.
    - Introduction of a bilingual terminology identifying feature in the Moses phrase tables. Automatic evaluation results show that the SMT quality can improve up to 7.8% over a system without such a feature.
    - Phrase pair filtering in the translation model creation process. Although being a highly experimental method, we have observed SMT quality improvements by up to 2.45% over a baseline system.

- Building in-domain and out-of-domain language models by using bilingual terminology and splitting monolingual corpora into in-domain and out-of-domain data sets. Our experiments have shown that monolingual corpora splitting using in-domain terminology increases SMT system quality by up to 7.34% over a baseline system.

The training level integration produces static SMT systems, which are useful for large translation tasks, however, translators often work on small projects for which training or even re-training of SMT systems is not economically justifiable. Therefore, in the TaaS project we have also investigated methods for dynamic integration of bilingual term glossaries in SMT systems during the translation. In order to support translation level integration, we have developed a translatable content pre-processing workflow that identifies terms in the source text, for morphologically richer languages generates possible inflected forms of terms and calculates their statistical probability. The pre-processed source text is then passed on to the SMT system in an XML format where terms with their translation equivalents are marked and the SMT system can select for each term the inflected form that best fits in a given context. Our automatic evaluation results show that dynamic integration of bilingual terminology in SMT systems improves the SMT quality by up to 26.9% over baseline systems.

To validate the automatic evaluation results, for the translatable content pre-processing workflow we performed manual comparative evaluation of baseline systems with the improved systems for three language pairs (English-Estonian/Latvian/Lithuanian). The results show that our methods provide considerable improvement in the SMT quality.

5. TaaS impact

The TaaS project has socio-economic impact in multiple areas.

TaaS **boosts the competitiveness of SMEs** in localisation and translation businesses by significantly decreasing the time spent on terminology work. Further productivity increase is achieved by the application of domain-adapted MT via the TaaS API to the terminology services. TaaS elaborates novel methods that **improve the quality of MT systems**. Novel approaches are elaborated and evaluated for terminology integration with SMT to adapt for domain specific translations showing quality improvement of up to 26.9%.

TaaS enables **new and efficient work patterns**: translators, terminologists, and language workers spend less time searching for term candidates and their translation equivalent candidates. TaaS increases the **competitiveness of European language technology providers** – the integration of terminology services into memoQ and OmegaT CAT tools and the TaaS API for the integration of terminology services with other applications provide a strong competitive advantage for CAT tool developers.

By providing innovative services for data acquisition from the Web, TaaS greatly increases the **availability of terminology resources** that are of high demand by language workers and are a prerequisite for efficient multilingual communication. TaaS provides the services for all official EU working languages, particularly advancing tools and resources for less-resourced languages thus contributing to the **crossing the language barriers**.

The TaaS terminology services **increase the usage of European language resources** – seeking term translations in the vast tangle of available terminology resources is beyond the workflow of a casual translator and/or terminologist and inefficient for a professional translator. TaaS provides efficient terminology services for the application of terminological data stored in IATE, EuroTermBank, TAUS Data, and other repositories.

Crossing the digital divide. Due to the growing volume of terms that appear in usage with increasing velocity, the lack of repositories of multilingual terminological data exacerbates the digital divide between larger and smaller EU languages. A wide accessibility to better and broader terminology now promotes the availability of information and learning and integration for representatives of smaller languages. Access to organised, available, and up-to-date terminology promotes the development and acceptance of appropriate new terms in a broader range of languages.

TaaS contributes to the **development and adoption of new WWW consortium standard** ITS 2.0 by enabling the ITS 2.0 application showcase that demonstrates new features for the multilingual Web.

TaaS **enriches the open European language resource infrastructure** META-SHARE providing for sharing about 3 million multilingual term candidates acquired from the Web (see Figure 27 and Figure 28).

TaaS services can become an essential **part of the European multilingual infrastructure** to serve the critical needs of the public sector and businesses for enabling language technologies.

Figure 27. TaaS services in META-SHARE

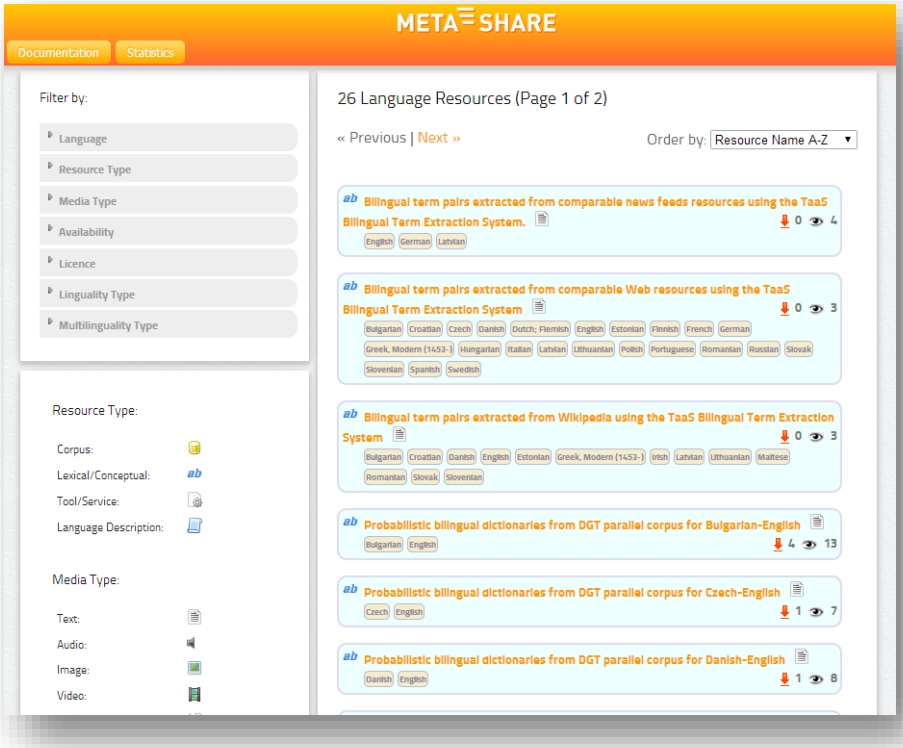


Figure 28. TaaS resources in META-SHARE

6. TaaS dissemination

TaaS dissemination raises awareness about the project and its results, makes the project visible, and promotes it. TaaS dissemination credo is to "get the right message to the right people in the right way at the right time" using the spoken word, written communication, and visual images as the three basic means of communication. The TaaS dissemination toolkit contains the project visual identity (logo, presentation template etc.), public website as the major public dissemination channel, communication collateral (introductory leaflet and poster, midterm leaflet and poster, final leaflet and poster), and 25 publications (the list of the TaaS publications in provided in section 6).

TaaS was promoted at 39 events with presentations and demonstrations to name just a few: Localization World, TAUS Asia Summit Beijing, the European Union Association of Translation Agencies event in Brussels, the MultilingualWeb workshop in Dublin, Rome, and Madrid, META-FORUM in Brussels, CHAT 2012 in Madrid, CHAT 2013 in Wiesbaden, and others (read news and see pictures on the project website). TaaS was endorsed via social and professional networks Facebook, Twitter, LinkedIn, Google+, ResearchGate, and others. A networking group on LinkedIn Terminology Services with more than 730 members, a professional topic Terminology Services on Scoop.it!, a Twitter account Terminology Services with more than 220 followers are dedicated to TaaS and are actively used to broadcast the news about TaaS and terminology in general.

6.1.1. TaaS Workshops

TaaS at CHAT 2012. TaaS was one of the three EU projects – co-organisers of "CHAT 2012: the Second Workshop on Creation, Harmonisation and Application of Terminology Resources" held on 22 June 2012 in Madrid, Spain (see Figure 29). The workshop was organised by Tilde and TaaS, ICT PSP project META-NORD (Baltic and Nordic Branch of the European Open Linguistic Infrastructure) and PF7 TTC project (Terminology Extraction, Translation Tools and Comparable Corpora).

The workshop programme was published on the workshop page and the workshop proceedings were published in Linköping Electronic Conference Proceedings, No. 72. CHAT 2012 continued a series of meetings that started as the first workshop CHAT 2011 in Riga, Latvia, in May 2011.

CHAT 2012 aimed at bringing together academic and business players in the terminology field and attracting holders of terminology resources. The workshop also focused on fostering the cooperation between EU projects and research and development activities in the area of terminology along with sharing experience and discussing recent advances of the consolidation, harmonisation, and distribution of terminology resources, as well as their practical application in various realms.

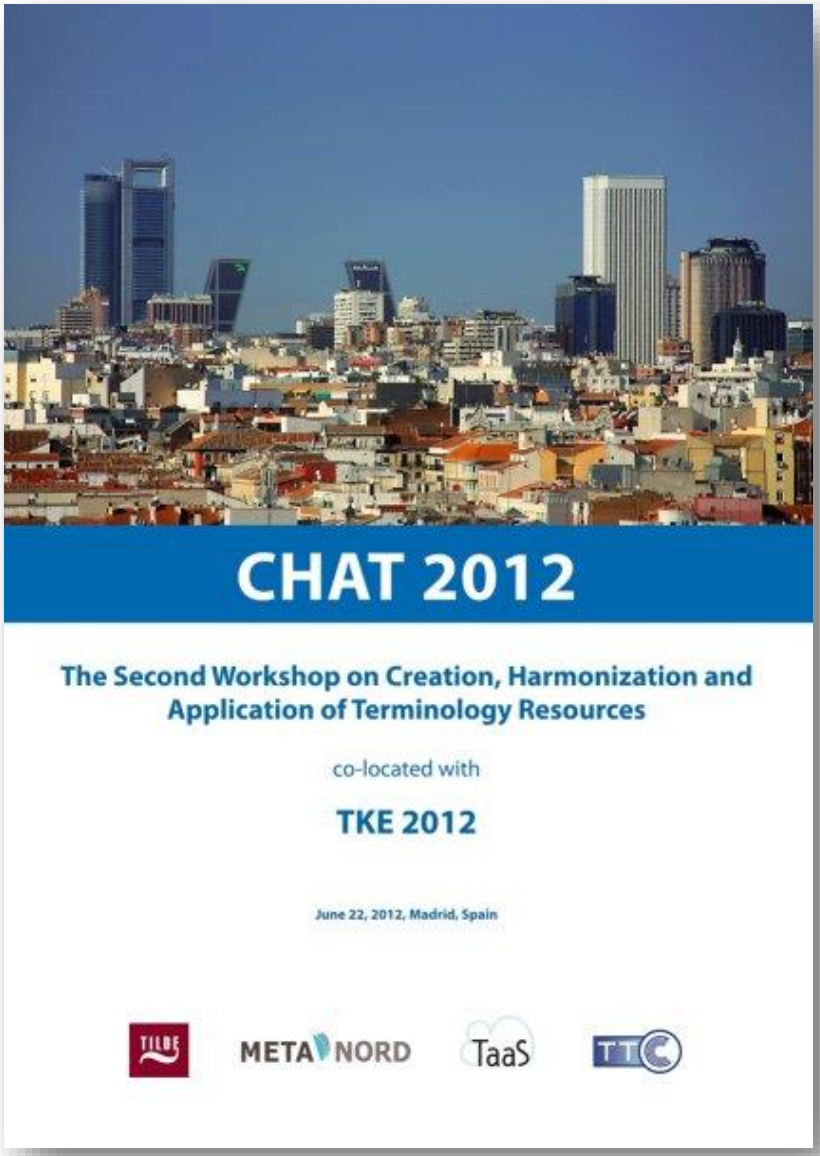


Figure 29. CHAT 2012 poster

TaaS at CHAT 2013. The TaaS consortium organised a full-day track "CHAT 2013: Creation, Harmonisation, and Application of Terminology" at the TCWorld/Tekom conference on the 7th November 2013, Rhein-Main-Hallen, Wiesbaden, Germany (see Figure 30). CHAT 2013 continued a series of meetings that began as CHAT 2011 in Riga, Latvia and followed by CHAT 2012 in Madrid, Spain.

The event in Wiesbaden was essential for the TaaS project as it brought together international terminology practitioners, business stakeholders, and researchers to discuss the latest advances and challenges in the terminology field and raise the profile of the TaaS initiative (see the TaaS first workshop public report).

Figure 30. CHAT 2013 Poster

TaaS Final Workshop. The second year of the project was fruitful for business and academia events. One of the outstanding events was the TaaS final workshop held on June 4, 2014 in Dublin at Localization World. The workshop programme was announced on the project website, the TAUS website, and other channels using promotion materials (see Figure 31).

Figure 31. TaaS Final Workshop Poster

Presentations covered various aspects of terminology and related trends: academic research, practical experience of LSPs, results of the EC-funded projects, and trend watching report (see Figure 32). Presentations are available on the project website and the TAUS website. Some presentations were published on slideshare.net under CC attribution rights. During the week upon the publication, presentations attracted high interest with 320 views.[12]

[12] For example, http://www.slideshare.net/TAUS/taas-workshop-2014-terminology-as-a-service-indra-samite-tilde and http://www.slideshare.net/TAUS/taas-workshop-2014-terminology-trends-firsthand-experience-as-a-blogger-maria-pia-montoro-intrasoft-international.

Figure 32. Presentations during the TaaS workshop in Dublin

6.2. Exploitation of the TaaS results

The TaaS terminology services have strong exploitation potential that will ensure their long-term sustainability and further advancement beyond the end of the EU co-funded period. The TaaS consortium partners will exploit the results of the project in multiple complimentary ways: terminology services will be provided for businesses, public sector, and for training and research in academia.

Tilde will operate the TaaS platform and maintain its services at least 24 months after the end of the project. TaaS services will be integrated with the LetsMT platform ensuring correct and consistent terminology in MT provided for eGovernments in Latvia and Lithuania as well as numerous other customers boosting the competitiveness and growth of Tilde's MT business. TaaS will be also used by Tilde localization team in localisation projects into Estonian, Latvian, and Lithuanian via the services for language workers provided on the TaaS platform. Multilingual and collaborative terminology services in language work will lead to more productive and efficient localisation business.

Kilgray released the latest version of its popular CAT tool memoQ 2014 on 10 June 2014 that provides a tight integration with TaaS. With all functionalities in memoQ, Kilgray works closely with users to see how this integration can better meet their needs.

TAUS will use the TaaS results to enhance the services offered through the TAUS Data platform to all business stakeholders. The new web crawlers integrated with TaaS will help TAUS to 'harvest' new data and support users with more languages and more domains to search and find terms and segments for their translation work. TAUS Data is also being used by organisations to train and customise MT engines. The increase of data volumes in new domains and language pairs will lead to improved performance of MT engines for many of the TAUS users.

IIM is based at Cologne University of Applied Sciences and TaaS data and project achievements will be used as a source for research in terminology related fields or as a resource for the creation of teaching materials. The functional specification developed during the project and containing information on technical and computer-linguistic requirements of terminology tools, theoretical background from the field of terminology, and on user needs will be useful in training terminology to give students an overview on the possibilities and requirements of a terminology tool. In addition, the project results will be useful as a source for scientific articles in relevant professional magazines. The TaaS domain classification elaborated by IIM and its mapping to the EuroVoc domain classification will be useful in following research projects or in teaching materials. Finally, argumentation papers and findings created during the discussion, development, and definition of the database model and architecture of the TaaS platform can be used as a resource for IIM teaching or research activities.

USFD will exploit various tools and data resulting in TaaS for the purpose of future scientific research and development. Different tools developed in the project will be exploited further to improve their qualities. USFD also plans to use the manual assessment data gathered at the evaluation task in TaaS for future research publications.

Finally, terminology services can greatly contribute to the creation of the European multilingual infrastructure to serve the critical needs of the public sector and businesses for enabling language technologies.

7. TaaS Publications

Ahmet Aker, Monica Paramita, and Robert Gaizauskas. 2013. Extracting Bilingual Terminologies from Comparable Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 402–411, Sofia, Bulgaria. Association for Computational Linguistics.

Ahmet Aker, Monica Lestari Paramita, Emma Barker, and Robert Gaizauskas. 2014a. Bootstrapping Term Extractors for Multiple Languages. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*, pages 483–489, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ahmet Aker, Mārcis Pinnis, Monica Lestari Paramita, and Robert Gaizauskas. 2014b. Bilingual dictionaries for all EU languages. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*, pages 2839–2845, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rahzeb Choudhury. 2013. Translation technology big data revolution. *MultiLingual Magazine*(January/February):39–41.

Tatiana Gornostay. 2014. Dreams of Better Terminology Tools. *Multilingual magazine*(April/May):44–45.

Tatiana Gornostay and andrejs Vasiļjevs. 2014. Terminology Resources and Terminology Work Benefit from Cloud Services. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1943–1948, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tatiana Gornostay and Andrejs Vasiļjevs. 2014. Cloud Terminology Services Facilitate Specialised Lexicography Work. In *Proceedings of EURALEX 2014*, Bolzano, Italy.

Tatiana Gornostay, Andrejs Vasiļjevs, and Roberts Rozis. 2012. Cloud-Based Infrastructure of Terminology Services and Resources. In *Proceedings of European Association for Terminology (EAFT 2012) Terminology Summit*, Oslo, Norway. European Association for Terminology.

Tatiana Gornostay, Olga Vodopiyanova, Andrejs Vasiļjevs, and Klaus-Dirk Schmitz. 2013. Cloud-Based Terminology Services for Acquiring, Sharing and Reusing Multilingual Terminology for Human and Machine Users. In *Proceedings of the TRALOGY II Conference "The quest for meaning: where are our weak points and what do we need?,"* Paris, France.

Iulianna van der (TermCoord) Lek. 2013. Interview with Tatiana Gornostay - Why Terminology is Your Passion? Interview with terminologists from all over the world.

Monica Lestari Paramita, Emma Barker, Ahmet Aker, and Robert Gaizauskas. 2014. Assigning Terms to Domains by Document Classification. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, Dublin, Ireland.

Mārcis Pinnis. 2013. Context Independent Term Mapper for European Languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*, pages 562–570, Hissar, Bulgaria.

Mārcis Pinnis, Tatiana Gornostay, Raivis Skadiņš, and Andrejs Vasiļjevs. 2013. Online Platform for Extracting, Managing, and Utilising Multilingual Terminology. In *Proceedings of the Third Biennial Conference on Electronic Lexicography, eLex 2013*, pages 122–131, Tallinn, Estonia. Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia) / Eesti Keele Instituut (Tallinn, Estonia).

Mārcis Pinnis and Raivis Skadiņš. 2012. MT Adaptation for Under-Resourced Domains – What Works and What Not. In Arvi Tavast, Kadri Muischnek, and Mare Koit, editors, *Human Language Technologies – The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012*, volume 247, pages 176–184, Tartu, Estonia, Estonia. IOS Press.

Inke Raupach. 2012a. Terminology as a Service (TaaS). *MDÜ journal*, 2012-4.

Inke Raupach. 2012b. Terminology as a Service (TaaS). *eDITion*, 2-2012:36.

Peter Reynolds. 2014. Using TaaS for Terminology Extraction.

Klaus-Dirk Schmitz. 2012. Terminologische Informationen und Dienste für Spracharbeiter. In *Tekom-Jahrestagung 2012*, pages 467–469, Wiesbaden, Germany.

Klaus-Dirk Schmitz and Tatiana Gornostay. 2013. Beyond the Conventional Terminology Work. In *The Third Workshop on Creation, Harmonization and Application of Terminology Resources (CHAT 2013)*, Wiesbaden, Germany.

Inguna Skadiņa, Andrejs Veisbergs, Andrejs Vasiļjevs, Tatjana Gornostaja, Iveta Keiša, and Alda Rudzīte. 2012. TaaS in the META-NET White Paper Series. In *META-NET White Paper Series "The Latvian Language in the Digital Age,"* pages 32, 73. Springer.

Raivis Skadiņš, Mārcis Pinnis, Tatiana Gornostay, and Andrejs Vasiļjevs. 2013. Application of Online Terminology Services in Statistical Machine Translation. In *Proceedings of the XIV Machine Translation Summit*, pages 281–286, Nice, France. The European Association for Machine Translation.

Matilda (TermCoord) Soare. 2013a. Terminology moves to "Cloud."

Matilda (TermCoord) Soare. 2013b. Why is Terminology your Passion? Interview with Prof. Dr. Klaus-Dirk Schmitz.

Andrejs Vasiļjevs, Mārcis Pinnis, and Tatiana Gornostay. 2014. Service model for semi-automatic generation of multilingual terminology resources. In *Proceedings of the 11th Conference on Terminology and Knowledge Engineering (TKE 2014)*, pages 67–76, Berlin, Germany.

Jost Zetzsche. 2014. Review of the TaaS platform. *Tool Box Journal (A computer journal for translation professionals)*(13-11-229).

8. Project Consortium

URL: http://www.tilde.eu

Project Coordinator
Tilde SIA
Vienibas gatve 75a
Riga, LV1004
Latvia

URL: http://www.fh-koeln.de/

Fachhochschule Köln
Fachhochschule Köln
Claudiusstr. 1
D-50678 Köln
Germany

URL: http://kilgray.com/

KILGRAY FORDITASTECHNOLOGIAI KFT
Beke Sugarut 72 2/3
Gyula, 5700
Hungary

URL: http://nlp.shef.ac.uk/

THE UNIVERSITY OF SHEFFIELD
Firth Court, Western Bank
Sheffield, S10 2TN
United Kingdom

URL: http://www.taus.net

TAUS B.V.
Oosteinde 9, De Rijp
1483AB
Netherlands