Cloud Terminology Services Facilitate Specialised Lexicography Work

Tatiana Gornostay and Andrejs Vasiļjevs
Tilde
E-mail: tatiana.gornostay@tilde.lv, andrejs@tilde.lv

Abstract

In this software demonstration paper we present an innovative cloud-based platform TaaS "Terminology as a Service" developed in an EU-funded project.[1] The TaaS platform provides language workers and language applications (human and machine users, accordingly) with the services to foster the creation, validation, harmonisation, sharing, and application of terminology resources. Under language workers we understand language professionals, for example, technical writers, editors and proof-readers, translators and localisers, terminologists and domain specialists, lexicographers and terminographers, and others. Under language applications (or machine users in other words), in the first place we consider computer-assisted translation (CAT) tools and machine translation (MT) systems (also, knowledge organisation systems in library and information science, search engines, and others). TaaS provides the following terminology services: terminology search in various sources, terminology identification in and extraction from user-uploaded documents, terminology visualisation in user-uploaded documents, translation equivalent lookup in and retrieval from various sources, terminology refinement and approval by users, terminology sharing with other users, collaborative working environment, and terminology reuse in other applications. Among other benefits for language workers, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based specialised terminology work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries.

Keywords: terminology service; terminology work; specialised lexicography

1 Introduction

Lexicography, as the theory and practice of dictionary development, is one of the most labour-intensive human activities. The creation of a new dictionary from the scratch and its delivery to an end user requires many resources in terms of time, labour, and finance. The main drawback of a conventional *paper* dictionary is its static and out-of-date content. In specialised lexicography, it is even more critical since terminology is developing rapidly along with its subject field and science in general.

To overcome the shortcomings of conventional lexicography, an electronic punch-card machine was first used to create a prototype of a modern electronic dictionary by Roberto Busa in the 20th century. His first work was based on the automatic linguistic analysis of the works of Saint Thomas Aquinas (he lemmatised the texts). Roberto Busa compared the invention of an "electronic book" (instead of a printing book) to the introduction of a printing book by Gutenberg (instead of a manuscript). Since that time automated lexicography has been developing boomingly.

Nowadays, with the evolution of information technologies, the Web, and data (for example, open data, linked data, free language resources etc.), the task of automated specialised lexicographic work is put in the first place. Routine processes have been delegated to a computer. An electronic, or computer-based, specialised dictionary is easy to update and manage, and its main advantage is its flexible, dynamic, and extensible (for example, in terms of new languages) character. Moreover, the new era of information technologies offers the new ways of dictionary representation, for example, on a tablet, mobile, and other devices, and the usage patterns of a dictionary (including a specialised dictionary) are changing with the course of time.

The integration of natural language processing tools within a lexicographer's working environment have made it possible for him/her to linguistically and semantically analyse and tag data and then to extract required pieces of information from it. In specialised lexicography it is possible to identify and extract term candidates automatically for further processing (for example, refinement, approval, sharing and reuse). Thus a lexicographer can consider hundreds thousands of terms in a certain subject field in comparison with that time when only several thousands (usually no more than 2000) could be included in a conventional specialised paper dictionary. This opportunity is critical particularly in emerging domains.

2 TaaS: Terminology as a Service

In this software demonstration paper we present an innovative platform TaaS "Terminology as a Service". The platform provides language workers and language applications (human and machine users, accordingly) with the services to foster the creation, validation, harmonisation, sharing, and application of terminological data. Among others, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-oriented, collaborative, multilingual, interoperable, portable, and cloud-based work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries.

1 The TaaS Beta was officially launched on November 1, 2013 and is publicly available for open Beta testing at https://demo.taas-project.eu.

TaaS is being developed within an industry-research collaborative project under the EU Seventh Framework Programme for Research and Technological Development.[2] The main objective of TaaS is to address the need for instant access to the most recent terms and for direct user involvement in the creation, harmonisation, and sharing of terminological data. The Beta version of TaaS was officially launched on November 1, 2013 and is publicly available for open Beta testing.
The concept of innovative cloud terminology services for language workers and language applications is presented in Figure 1 below.

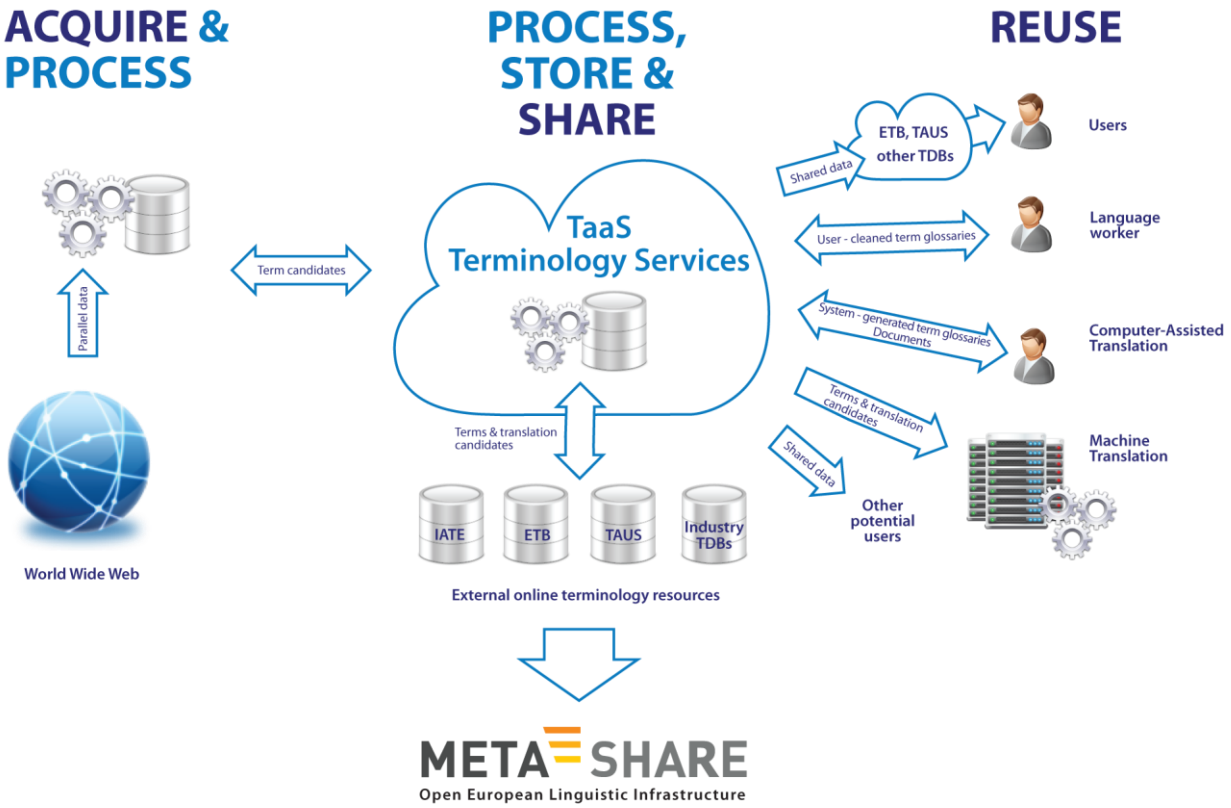

Figure 1: TaaS innovative cloud terminology services for language workers and language applications.

TaaS provides user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based terminology services to perform the following tasks:

- Search terminology in various sources;
- Identify term candidates in user-uploaded documents and extract them automatically applying linguistic and statistical processing;
- Visualise term candidates in user-uploaded documents;
- Look up translation equivalent candidates in various sources (for example, existing external terminology resources EuroTermBank[3], IATE[4], TAUS Data[5], and possible other resources, as well as automatically extracted bilingual terminological data stored in the TaaS Shared Term Repository and used as an additional internal source for target translation lookup);
- Refine term candidates and their translation equivalent candidates;
- Approve refined terminology;
- Share terminology with other users;
- Collaborate with colleagues in user-friendly working environment;
- Use terminology in other applications via TermBase eXchange ISO-standardised format (TBX), tab-separated value (TSV), and comma-separated value (CSV) export formats and via the TaaS Application Program Interface (API).[6]

To perform most of his/her work in TaaS, the user has to sign up for the services. However, for its non-signed users, TaaS provides the service for terminology search in two sources – TaaS database, which consists of TaaS users' terminology collections made public by its users, and EuroTermBank, which is the largest European online term bank, providing

[2] The reference to the project will be added in the full paper. The detailed information about the project has been omitted for the purpose of the blind review process.
[3] www.eurotermbank.com
[4] http://iate.europa.eu
[5] www.tausdata.org
[6] For TaaS API contact TaaS team via e-mail langserv@tilde.com.

access to more than 2 million standardised terms from more than 100 national terminology resources in 27 languages. For advanced search, the user has to select the source and target language, domain (a.k.a. subject field), and the source to be searched in (see Figure 2).

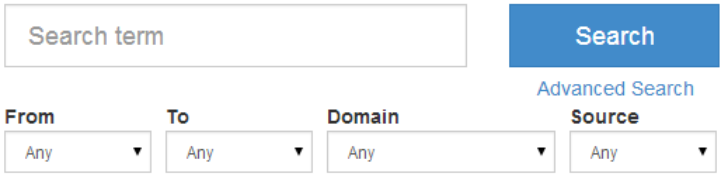

Figure 2: Search form in TaaS.

For signed users the work in TaaS is organised in projects. A signed user gains access to full TaaS services. To start his/her work, the user has to create a new project indicating the source and target language and the domain the user works in (it is relevant to user document(s) domain). More than 10 input format for user documents are supported. The user might also want to specify optional properties, such as product, customer, project description, and the business unit (in case of a corporate user) (see Figure 3).

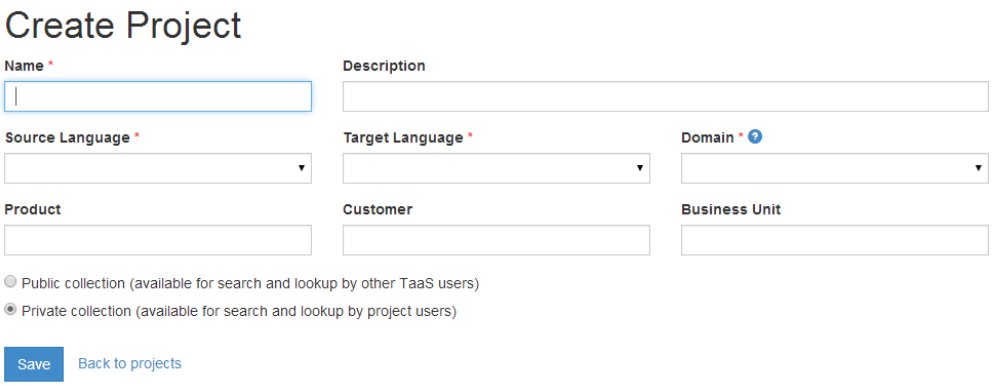

Figure 3: Creation of a new project in TaaS.

TaaS also provides a default project with project properties already set for demonstration purposes. Finally, the user has to set the status of his/her project – private or public. If the status of a project is public, the user's approved terminology will be available for search and lookup by other TaaS users; if the status of a project is private, the user's approved terminology will be available only to project users.

The user can start his/her work with TaaS by using the default project or creating a new project. In both cases, the user is an administrator of his/her project.

TaaS provides facilities for project sharing among users if they work in a team. This functionality that typically involves an interchange of non-confidential, non-competing, and non-differentiating terminology across various actors is highly rated by users. Recent surveys have shown that up to 60% of terminology resource users would share their resources with the community. The concept of sharing, unfortunately, is not present in the current management of major terminology databases and term banks. Instead of providing the opportunity for users to contribute their data, major term banks typically keep to the traditional one-way communication of their high-quality pre-selected terminological data.

To share his/her project with other users, the user has to add their e-mails and assign their roles. There are three available roles to a new user of the shared project: administrator, with full access rights; editor, with limited access to editing rights; and reader, with limited access to reading rights. One project can have more than one administrator; however, the owner of the project (the user, who has created the project) must consider assigning the administrator's role to other users of his/her project as these users will get full access, including the right to delete the project and its terminology collection. The Administrator's role is usually assigned to the project manager in the translation team, who adds documents to the project, and these are later processed by a terminologist, translator(s), editor(s), and other translation team members (see Figure 4).

The main usage scenario for the TaaS services is when the user uploads his/her document(s) under the created project, in order to then execute the terminology processing. TaaS supports user document upload in more than 10 formats including the most widely used MS Word, Excel, and Power Point formats as well as the Portable Document Format (PDF), the XML Localisation Interchange File Format (XLIFF), and others. The open Beta version has certain limitations in terms of file and project size.

The terminology extraction service performs automatic extraction of monolingual term candidates from user- uploaded documents using generic or language specific terminology extraction techniques.

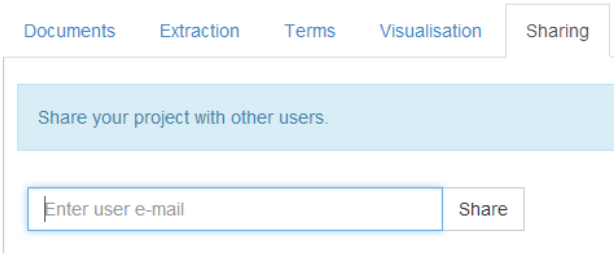

Figure 4: Project sharing in TaaS.

The user can customise the terminology extraction process. He/she can select one or more available (on the platform) terminology extraction tools for term candidate identification in user-uploaded documents. There are two term identification tools integrated into TaaS at the moment. These are the Tilde Wrapper System for CollTerm (TWSC)7 that includes language specific patterns and morphological analysis and Kilgray Term Extractor that applies generic statistical approach to all supported languages. It is recommended to select the first tool; however, the statistical tools might also be of help in certain cases, for example, when linguistic processing produces insufficient results.
The platform provides the service for automatic retrieval of translation equivalents (for the extracted monolingual term candidates) in user-defined target language from different public and industry terminology databases.
The following terminology resources are available for translation equivalent lookup for term candidates identified in user-uploaded documents:

- TaaS public collections shared by other TaaS users;
- Terminology collections owned by the user;
- EuroTermBank;
- Inter-Active Terminology for Europe (IATE), an inter-institutional terminology database of the European Union[8];
- TAUS Data that stores shared translation memories;

TaaS database of raw bilingual terminological data automatically extracted from original and translated texts (a.k.a. comparable and parallel corpora) on the Web.
In the dynamic pace of technological developments and societal changes, new terms are coined every day by industry, translation and/or localisation agencies, collective and individual authors. Although these terms can be found in different online and offline publications, the inclusion of new terms in online public terminology databases and term banks takes months or even years, if it happens at all. As a result, terminology databases and term banks 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 public terminology resources.
At the same time many new terms and their translations can be found on the Web – in multilingual websites, online documents, support pages, etc. TaaS provides four bilingual terminology extraction workflows for Web data: one workflow for terminology extraction from parallel data and three workflows from comparable data. The latter three are customised to collect terms from comparable news corpora, from multilingual Wikipedia, and from focused comparable corpora, respectively.
Web data are collected and then automatically processed. As a result, a list of bilingual raw term candidate pairs are extracted and fed into the TaaS terminology repository. During the execution of a terminology project at the translation candidate lookup step, these data are retrieved and proposed to the user for his/her validation. Thus the TaaS aligns the speed of terminology resource acquisition with the speed at which the content is created by mining new terms directly from the Web.
The data collection process is ongoing constantly feeding the TaaS repository with new terms. By April 2014, the TaaS database included more than 8 M bilingual term pairs extracted from the Web data.
TaaS provides facilities for cleaning up raw terminological data extracted automatically that is noisy and needs validation by users. The process of validation can be regarded as a three-step procedure:

- monolingual validation (deletion of "unwanted" and/or unreliable term candidates, definition of termhood, term variant identification, deduplication, deletion of "incorrect" extraction, for example, a part of a longer noun group, synonym identification etc.);
- bilingual validation (bilingual checking of term candidates and their translation candidates, defining the right translation for the source term, deletion of irrelevant and/or incorrect translations, etc.);
- validation in context.

As soon as extraction finishes, the user can see extracted terms from his/her documents and their translation equivalents retrieved by TaaS. The user can hover over terms to get additional information, such as grammar, source, and context (see Figure 5).

[7] See the ACCURAT Toolkit 3.0 at www.accurat-project.eu.
[8] http://iate.europa.eu/

Figure 5: Clean-up and validation of raw terminological data in TaaS.

The user can approve terms with a single click and add translations him-/herself, is the right translation from proposed translation candidates is not found.
An extracted term with its translation equivalent(s) forms a terminology entry. For advanced purposes, the user might want to edit a term entry in full entry view using the term entry editor and to add additional information about terms, for example, definitions, notes, grammatical information, and usage properties, such as term type, register, administrative status, temporal qualifier, geographical usage, and frequency. The history of editing is saved and is seen in the full entry view. The user might also want to see term candidates identified by TaaS in his/her documents, and the visualisation functionality is available for this purpose (see Figure 6).

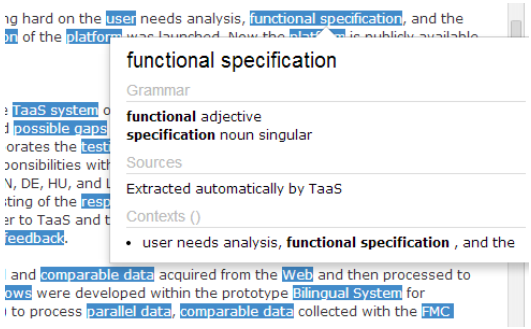

Figure 6: Visualisation of term candidates in the user's document in TaaS.

Validated terminological data can be exported and then reused in other working environments.
During the analysis of user needs and requirements, we also proved our hypothesis that terminology, as a language resource, is central for the second large group of users – language applications (the first user group is represented by language workers). We have already performed first successful experiments on the integration of terminological data acquired within TaaS into the statistical MT system. At the time, the memoQ CAT tool[9] owned and developed by Kilgray, the TaaS project partner, is being integrated with TaaS via the TaaS API developed in the project and available for machine users.
TaaS demonstrates the efficacy of its terminology services within the following practical usage scenarios:

- For language workers: simplification of processing, storage, sharing, and application of task-specific multilingual terminology.
- For computer-assisted translation (CAT) tools: instant access to term candidates and translation equivalent candidates via the TaaS API.
- For statistical machine translation (SMT) systems: support for domain adaptation by a dynamic integration with TaaS-provided terminological data via the TaaS API.

At EURALEX the usage scenario for language workers with the emphasis on specialised lexicography work will be demonstrated online.[10]

3 Conclusion

In this software demonstration paper we have presented an innovative cloud-based platform TaaS "Terminology as a Service" developed in an EU-funded project. At the present time, TaaS is a unique dynamic cloud-based solution that provides a wide range of terminology services. We foresee the potential of the established platform for a wide range of user groups, both language workers and language applications. Among other benefits for language workers, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based specialised terminology work. TaaS fills the gap of innovative environment to

[9] See the description at http://kilgray.com/products/memoq.
[10] Live demonstration requires Internet access.

speed up the development of specialised dictionaries This opportunity is critical particularly in emerging domains. At the EURALEX Congress the platform is demonstrated in real-time during the three days of the event.

Acknowledgements

Research within the TaaS project, leading to these results, has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 296312.

The TaaS platform is a result of fruitful collaborative work of the project partners – coordinator and lead developer Tilde (Latvia), research partners Cologne University of Applied Sciences (Germany) and University of Sheffield (UK), industry partners Kilgray (Hungary) and TAUS (Netherlands).