TaaS 1st Period Publishable Summary

Project Motivation

With the evolution of the Internet and cloud computing, there has been a shift towards collaborative solutions with a focus on user/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/document life cycle from creation, translation, localisation, publication, and numerous other information management steps to ensure efficient and precise communication.

The following tasks have attracted the most interest recently: robust automatic extraction of multilingual terminology; development, administration, and integration of online terminology resources; utilisation of multilingual terminology resources in language technology applications; application of terminology resources in machine translation; collaborative terminology management and sharing; consistency assurance of corporate and industry multilingual terminology; interoperability and harmonisation of terminology resources.

The static model which supports conventional terminology work cannot keep up with the growing demand of the volume and management of information. The motivation for the TaaS project[1] 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.

Project Main Objective

TaaS addresses a wide range of the abovementioned challenging scientific and technological tasks. The main **objective** of the TaaS project is to establish a cloud-based platform for acquiring, processing, and reusing multilingual terminological data.

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

Project Main Result

The main **result** of the project will be the innovative platform TaaS "Terminology as a Service" for acquiring raw terminological data, cleaning up these data, and then, sharing and reusing terminological data, based on cloud computing. TaaS offers the following cloud-based terminology services (see also Figure 1):

- **Import** of files in different formats widely exploited by users, e.g., DOC(X), PDF, XML-based formats like XLIFF, etc.;
- Automated **extraction** and **retrieval** of monolingual term candidates (from documents uploaded by users) using state-of-the-art linguistically and statistically motivated terminology extraction techniques;
- Automatic **lookup** for term translation equivalents (for monolingual term candidates automatically extracted from documents uploaded by users) from the largest publicly available terminology databases, such as IATE and EuroTermBank, as well as statistical terminological data acquired from publicly available parallel and comparable Web data by use of state-of-the-art linguistically and statistically motivated terminology extraction and bilingual terminology alignment techniques;
- **Creation of monolingual and bilingual terminology** collections in user-defined languages;
- **Collaborative terminology clean-up**, e.g., deletion of irrelevant or unreliable term candidates and "incorrect" extraction (e.g., 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 resulting terminological data with major terminology databases and banks;
- **Reuse** of terminology collections in various applications within different human and machine usage scenarios via the TaaS application user interface (API) and export of files in different formats widely exploited by users, e.g., TSV, CSV, and TBX.

Project Usage Scenarios

TaaS demonstrates the efficacy of its terminology services within the following usage scenarios:

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

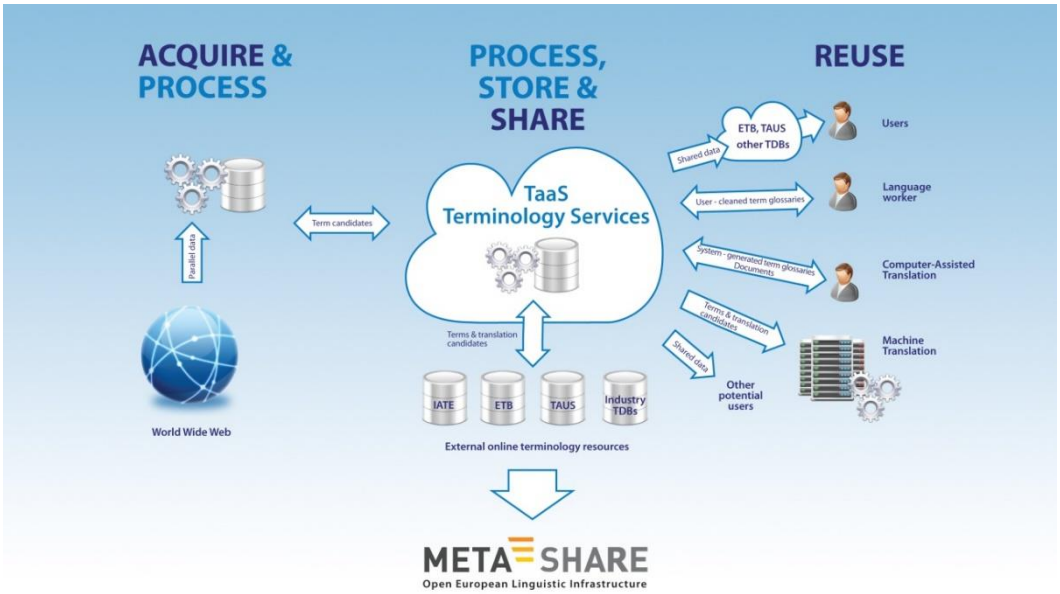

Figure 1. General schema of TaaS terminology services

Project Results for the 1st Period of the Project

During the first period of the project, the Consortium undertook a wide range of tasks that include methods and techniques for target user group, user needs, usage scenario, and market analysis, software programme testing and evaluation, parallel and comparable data acquisition from the Web, user-provided file filtering, extraction and retrieval of monolingual term candidates, bilingual translation equivalent lookup, bilingual term alignment, technical analysis of TaaS requirement, TaaS cloud-based infrastructure setup, TaaS facilities development, TaaS integration with machine users, and other tasks.

The following results were achieved during the first period of the project:

- **TaaS target user groups** (i.e., potential users of TaaS terminology services), their **terminological needs** and typical **usage scenarios** of terminology services were analysed on basis of an online TaaS User Needs Survey (with appr. 1,800 respondents). The following target user groups were identified: language workers (human users) and computer-assisted tools and statistical machine translation systems (as representatives of machine users).
- The **prototype bilingual term extraction system** was developed including the three workflows: parallel and comparable Web resource collection, monolingual term candidate extraction, and bilingual term candidate alignment, on basis of the review and evaluation of best practices and state-of-the-art for the acquisition and processing of data from the Web, terminology extraction and alignment methods, techniques, and available tools.
- The **TaaS Shared Term Repository** was designed, developed, and populated with raw terminological data (i.e., raw aligned term pairs) in four domains and eight languages extracted from parallel and comparable corpora acquired from the Web. Data model for the TaaS Shared Term Repository was proposed on basis of the current status of relevant international standards for data exchange and data modelling. The TaaS Shared Term

Repository provides an online access point for the **presentation layer** (human and machine users of TaaS terminology services), the **logic layer** (TaaS terminology services, user management, and others), and the **data layer** (storage of data at different levels, e.g., user data, project data, and terminology).

- The preview version of the **TaaS platform** was developed, tested, and then deployed externally and run publicly available on the Amazon Cloud Service.[2] TaaS interdependent components were integrated, e.g., internal databases, import file filtering, terminology service infrastructure, terminology processing modules, external services and terminology resources, and others. The terminology extraction process was created on the High Computing Cluster.

Project Dissemination

TaaS dissemination raises awareness of the project and its results and makes the project visible, as well as promotes it. The 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 1st period of the project laid the foundations of the TaaS dissemination toolkit: the project website, communication collateral, branding etc. for the exploitation to be performed during the 2nd period of the project and beyond it. The project website www.taas-project.eu was established to provide the basic public dissemination channel. Initial TaaS leaflet and poster were designed, published, and distributed among partners. TaaS was promoted at 13 events with presentations and demonstrations (to name just a few, Localization World Singapore, TAUS Asia Summit Beijing, the European Union Association of Translation Agencies event in Brussels, the MultilingualWeb workshop in Dublin and Rome, META-FORUM in Brussels, the second workshop CHAT 2012 in Madrid, Tekom events, and others).

TaaS was widely promoted via social and professional networks, e.g., Facebook, Twitter, LinkedIn, Google+, ResearchGate, and others. A professional group was dedicated to TaaS activities on LinkedIn (with more than 555 participants)[3] and a professional topic "Terminology Services" was initiated on Scoop.it!.

The third workshop CHAT 2013: Creation, Harmonisation and Application of Terminology Resources was announced to be held at the Tekom/TCWorld Conference and Fair in Wiesbaden in November 2013, and the TaaS project is the main organiser. TaaS prepared and submitted two presentation proposals to CHAT 2013: "Beyond the conventional terminology work" and "Welcome to the cloud! Terminology as a service" (both accepted).[4]

Project partners also prepared 10 publications directly discussing TaaS motivation and concept or mentioning the TaaS initiative.

[2] The Preview version of the TaaS platform is available at: www.demo.taas-project.eu.

[3] http://www.linkedin.com/groups/TTC-Terminology-Extraction-Translation-Tools-3710659

[4] The event programme was already published on the TaaS website: http://www.taas-project.eu/index.php?page=alias