Cloud-Based Terminology Services for Acquiring, Sharing and Reusing Multilingual Terminology for Human and Machine Users. Session 4 - Terminology and Lexicology

Tatiana Gornostay, Olga Vodopiyanova, Andrejs Vasiljevs, Klaus-Dirk Schmitz

TRALOGY

Cloud-Based Terminology Services for Acquiring, Sharing and Reusing Multilingual Terminology for Human and Machine Users

Tatiana Gornostay
Tilde, Latvia
tatiana.gornostay@tilde.lv

Olga Vodopiyanova
Institute for Translation and Multilingual Communication, Cologne University of Applied Sciences, Germany
olga.vodopyanova@fh-koeln.de

Andrejs Vasiļjevs
Tilde, Latvia
Andrejs@Tilde.lv

Klaus-Dirk Schmitz
Institute for Translation and Multilingual Communication, Cologne University of Applied Sciences, Germany
klaus.schmitz@fh-koeln.de

TRALOGY II - Session 4
Date d'intervention : 18/01/2013

This paper presents the concept of the cloud-based terminology services for acquiring, sharing and reusing of multilingual terminology for human and machine users. An ongoing "Terminology as a Service" project was initiated to establish the TaaS platform addressing user needs and providing online core terminology services for key terminology tasks. The paper describes the main target user groups of the platform. The problems that language workers (technical writers, translators, interpreters, terminologists, editors etc.) encounter when working with terminology are analysed on basis of the results of the survey performed within the project.

lien video : http://webcast.in2p3.fr/videos-cloud_based_terminology_services

Introduction

Terminology is a spine of the professional communication and of a document within its life cycle, including the creation, publication, translation, localisation and other document management processes. Furthermore, terminology is of vital importance for brand consistency and customer satisfaction within businesses. Moreover, in the context of the multilingual Europe, the role of terminology and is undoubtedly even more important than ever to insure that people communicate efficiently and precisely.

Terminology is developing rapidly and every day the volume of terminological data grows along with the explosion of information available on the web. Recent surveys show that the situation in terminology management has not experienced significant changes since past decade. Still there are evident needs for an instant access to the most up-to- date terms, collaborative models for the acquisition, processing and sharing of multilingual terminology and facilities for the reusing of terminology resources in various applications within different usage scenarios.

Translators, editors, technical writers and other language workers spend up to 30% of their working time on terminology research. In some cases terminology research can consume more than 30% of overall working time, for example, in the translation of technical specifications [Massion, F. (2007)]. A language worker usually needs immediate answers to his/her terminology requests but due to time and cost constraints proper terminology search is often skipped. Resulting errors in term usage affect not only translation/localisation productivity and overall costs but also influence further stages of documentation life cycle, for example, failures in product technical support, client request processing etc.

1. Terminology as a Service

The current static models for core terminology tasks cannot keep up with the increasing demand. The concept of sharing, unfortunately, is not really present in the current management of major terminology resources either. 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 resources.

The core objective of the "Terminology as a Service" project[1] 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 project is developing an innovative online platform TaaS for the acquisition, sharing, and reuse of multilingual terminology and keep it up-to-date on a continuous basis by involving users in terminological data clean-up. TaaS will provide the following online core terminology services for key terminology tasks:

- Automatic extraction of monolingual term candidates using the state-of-the-art terminology extraction techniques from the documents uploaded by users;
- Automatic recognition of translation equivalents for the extracted terms in user-defined target language(s) from different public and industry terminology resources (e.g., TAUS, IATE, EuroTermBank and others);
- Automatic acquisition of translation equivalents for terms not found in existing terminology resources from parallel and/or comparable web data using the state- of-the-art terminology extraction and bilingual terminology alignment methods;
- Facilities for the platform users for cleaning up (i.e., editing, deleting) of automatically acquired terminology;
- Facilities for terminology sharing and reusing: APIs and export tools for sharing the resulting terminological data with major term banks and reuse in various applications within different usage scenarios.

(1) www.taas-project.eu

Multilingual consolidated and harmonized terminology is already utilised as data in the process of human translation and now it is also being developed as a web-based service for machines as users, for example, machine translation systems, indexing systems, search engines etc. This has the potential to vastly enhance the quality of language tools and natural language processing in general.

The cloud-based platform for terminology services TaaS will demonstrate the efficacy of reusing acquired and user-cleaned terminology resources within the following usage scenarios:

- Simplify the process for language workers to prepare, store and share of task- specific multilingual term glossaries.
- Provide an instant access to term translation equivalents and translation candidates for professional translators through computer-assisted translation (CAT) tools.
- Adapt a statistical machine translation (SMT) system to a certain domain by dynamic integration with terminological data.

Thus, TaaS addresses the terminology needs for both human (language workers) and machine users, for example, CAT tools, machine translation (MT) systems, authoring and content management systems, web crawlers and information retrieval systems (for example, cross-lingual information retrieval) which exploit term lists as seeds for acquiring web data etc. (see Figure 1).

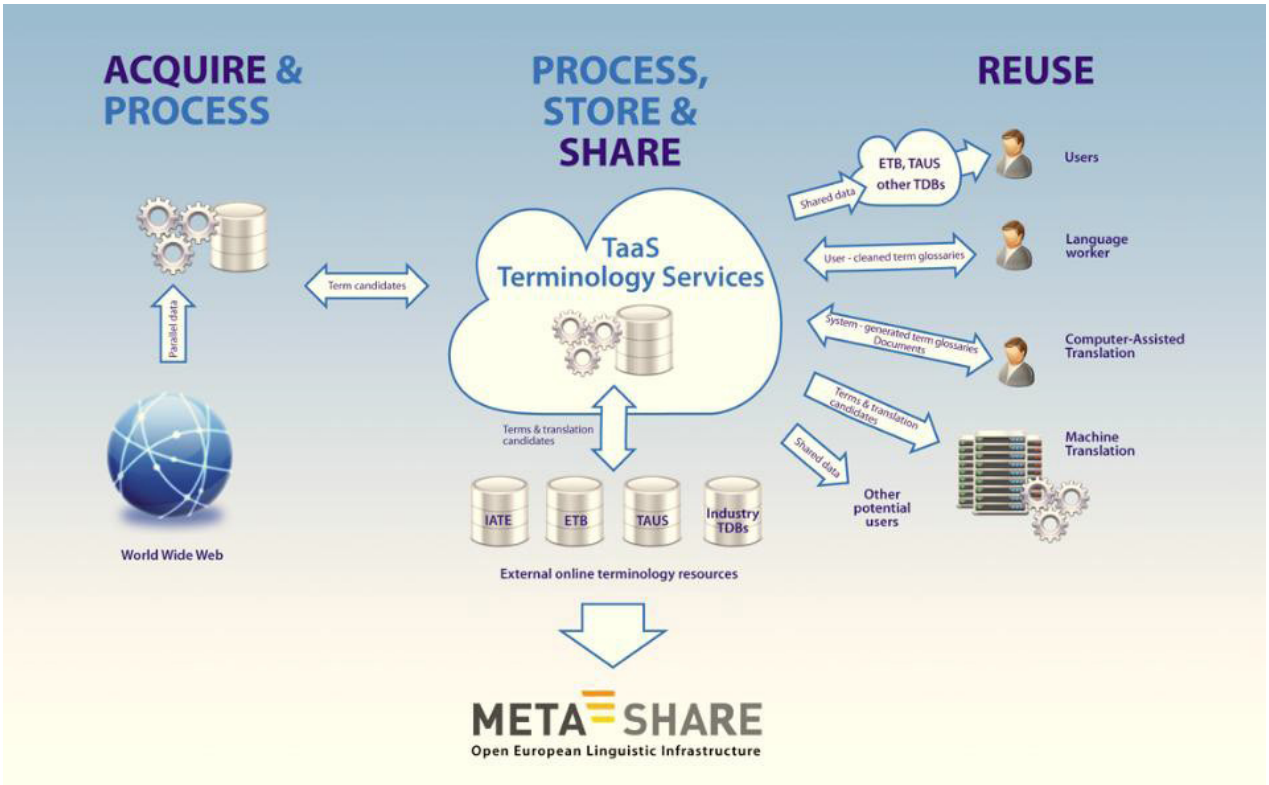

Figure 1. TaaS Concept

Machine users also need terminological data in order to improve the translation quality of specialised texts and require different additional term-specific information for automatic processing than human users. The project analyses the needs of machine users by studying several sample systems to identify the type, structure and format of terminological data needed

by this kind of users. Integration of TaaS services for machines as users of shared terminological data will be enabled via elaborated API.

Overall, the TaaS strategy of open access to data, API access to resources and services and reliance on terminological data exchange standards opens a variety of applications and use cases.

2. Target User Groups of TaaS

One of the potential user groups, probably the biggest one, comprises human users (as opposed machine users, for example, CAT tools and MT systems). The first step of the project was intended to define the potential (human) user groups of the TaaS platform. The second one aimed at finding out the terminological needs of the potential users and identifying most frequent problems they face when doing online terminology search.

Terminology work deals with subject field specific vocabulary and related additional terminological information. Consequently, all groups of professions which focus on collecting, processing and using of these kinds of data could be identified as potential user groups of TaaS.

Technical writers/editors can be named as the first group as the members of this professional field use terminology "first-hand" in order to produce subject-specific texts in their field of knowledge.

As terminology cannot be treated as isolated monolingual data, target language equivalents of terms should be found in order to correctly transfer (translate) subject field specific texts from one language into another. In this context, translators, interpreters and software localizers can be named. Interpreters "mediate" the respective contents in oral way and the professional activity of translators is the written transfer of subject filed specific texts. Similarly to translators, software localizers adapt software and IT texts to target markets using the appropriate terminology.

Proceeding with the potential users of TaaS, there are certain professional groups which have to master the terminology of a given subject field and to regularly enrich their knowledge in accordance with new developments. In this regard, language teachers and domain experts were identified.

Finally, terminologists have to be mentioned as a potential user group since their professional activities include collection, systematization, evaluation and definition of terminology, as well as terminology maintenance.

3. Terminological needs of potential TaaS users

Having identified the potential user groups of TaaS, the project was eager to find out what the users really need when using online terminology services. For this purpose, a respective questionnaire was developed and an online survey was conducted.

3.1 Planning the survey

The professional groups of technical writers/editors, translators, interpreters, software localizers and terminologists as well as language teachers/learners and domain experts were defined as potential users of the TaaS platform. The respective multipliers for spreading the questionnaire were identified. These are associations of translators, interpreters, technical writers and terminologists which contribute to the exchange of information, experiences and news in the field of terminology work (for example, tekom - German Professional Association for Technical Communication, FIT Europe – regional centre Europe of the International Federation of Translators, BDÜ – German Federal Association of Interpreters and Translators, DTT – German Association for Terminology etc.).

A questionnaire was designed for the survey which asked for the professional background of the interviewed persons, the types of terminology work and terminology management they exercise as well as the problems they usually face when doing terminology work. 1,782 persons participated in the survey which ensured a reliable average for the analysis, i. e., representative survey results.

3.2 Results of the User Needs Survey

The first questions were intended to state the professional background of the interviewees. As the survey results showed, the main professional groups were represented by technical writers and translators (see Figure 2).

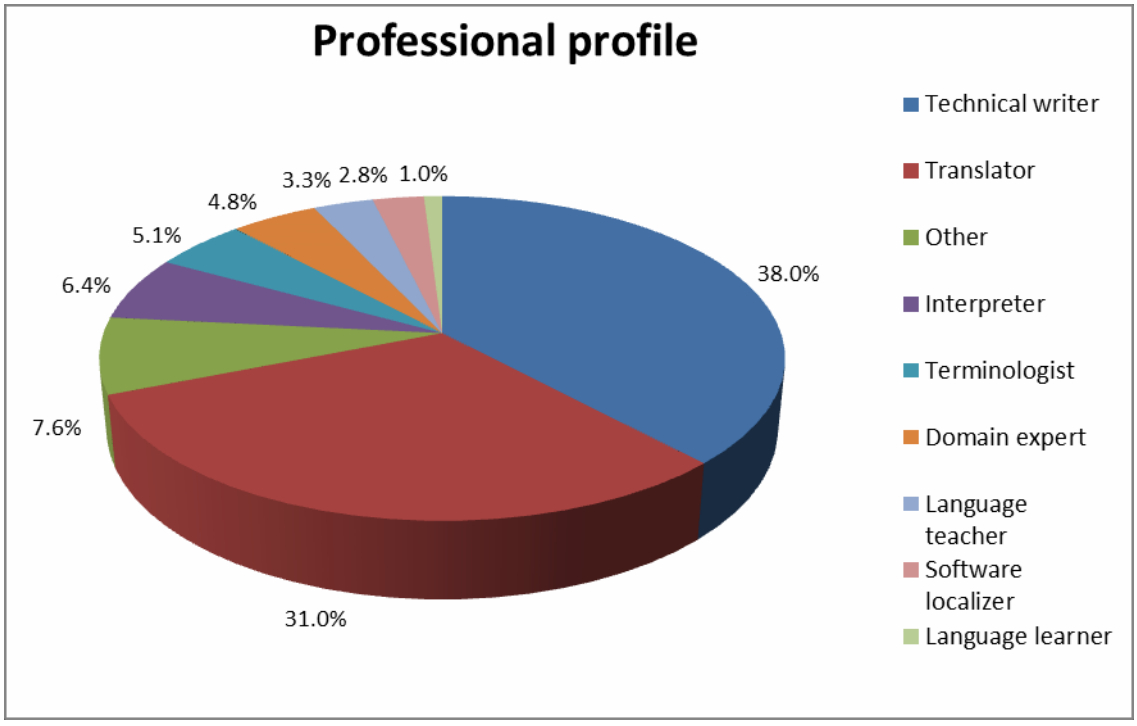

Figure 2. Professional profile

The prevailing working status of the potential users of TaaS are "Employed with an industrial company" with 51.1% (mainly technical editors/writers and terminologists) and "Freelance" with 34.6% (mainly translators and interpreters).

Subsequently, it was of interest to identify where and how the interviewed persons search for terminology and the related information to consider the importance of terminology services offered online. The survey results showed that keyword search with search engines, online terminology data bases, online encyclopaedias and web-based translation memories are the most popular ways to search for terminology.

Concerning the technical requirements to the TaaS platform, it was also crucial to find out what programs and formats the participants mainly use when editing, translating and managing terminology. This is important for the formats and interfaces TaaS should provide. TXT, DOC, XLS, TMX, and XLIFF were the prevailing formats which consequently were shortlisted for the TaaS platform.

Concerning terminology work in general, 83.4% of the interviewed persons consider terminology work to be very important or quite important. Nevertheless, the time the participants spend on terminology work within their regular professional activities was estimated at less than expected (see Figure 3).

It was also important to identify the kinds of terminological information that is mainly searched for by the potential TaaS users and the subject fields with the greatest demand on terminology. Both will be used to specify the type of terminological information and the domains the TaaS platform will

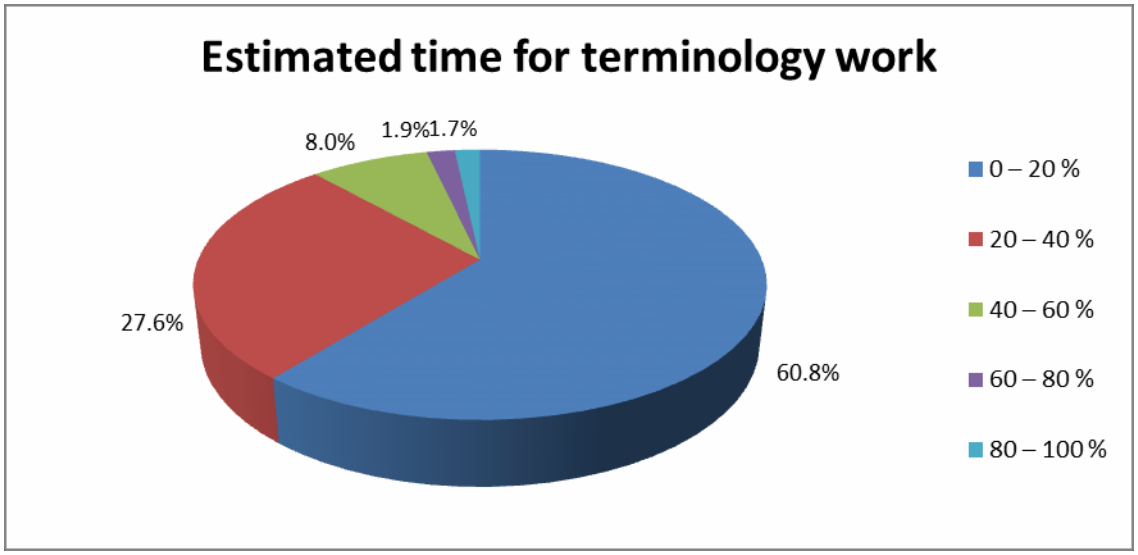

Figure 3. Estimated time for terminology work

provide. As the results in Figure 4 show, the participants are mostly interested in finding terms in the source language (SL) and target language (TL) as well as definitions in both languages.

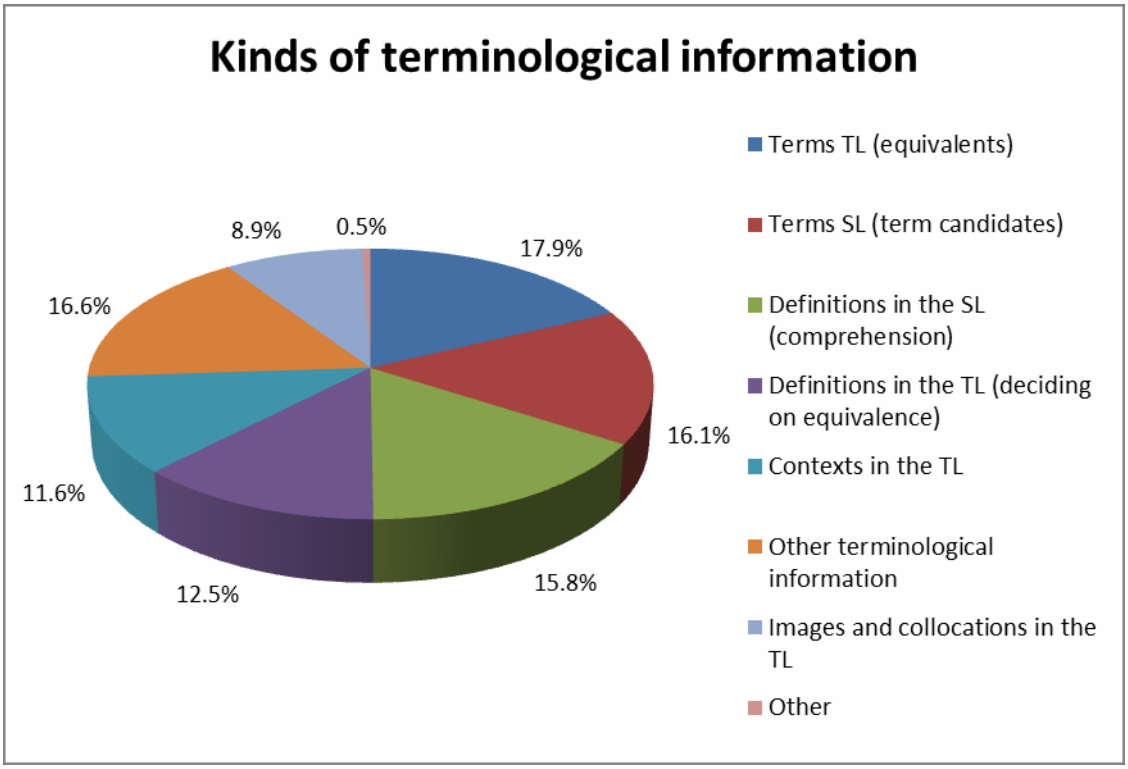

Figure 4. Kinds of terminological information

As far as the subject fields are concerned, mechanical engineering, information technology, electrical engineering, law, economics, and medicine were prioritised by the interviewees.

Subsequently, the survey was intended to state what problems the survey participants face while searching for terminological data. The results are shown in Table 1.

Table 1. Problems while searching for terminology

Problematic issues	%
Lack of resources/Insufficient terminology management	24.5
Poor quality/Up-to-dateness	14.5
Lack of information (contexts, definitions, local and company-specific use etc.)	13.6
Lack of convincing verification/Misleading information online	12.3
Inconsistent use of terminology	10.9
Technical issues	8.9
Lack of covered specific subject fields	6.6
Lack of freely accessible standardized data/ Unwillingness to share information	5.4
Ambiguity/Synonymy	4.5
Lack of covered languages and language pairs	4.1
Lack of equivalents	2.4
Other	11.2

The interviewees were also asked to express their opinion on the optimization potential in the area of online terminology search (see Table 2).

Table 2. Optimisation potential

Optimisation potential	%
Database scale (subject fields, languages, terms)	17.6
Quality assurance (validation of terms and sources, up-to-dateness)	15.4
Database content (data categories like context, definition, source etc.)	9.6
Pooling of related sources (links to other online terminology resources)	8.4
Collaboration (capability for discussion, feedback, joint elaboration of definitions etc.)	8.3
Technical issues (easy and fast to operate, interfaces to differentprograms, upload/download in different formats etc.)	7.5
Search/filter functions (e.g. phrase search, phrase + subject search etc.)	5.8
Other	32.2

Finally, the participants were asked if they were willing and interested to share their own terminology resources and if yes, if there were any preconditions. This information is also very important, for example, in terms of access rights, quality assurance but also in terms of the technical specification of TaaS. 60.5% of the interviewees answered in the affirmative. The most frequently mentioned preconditions are shown in Figure 5.

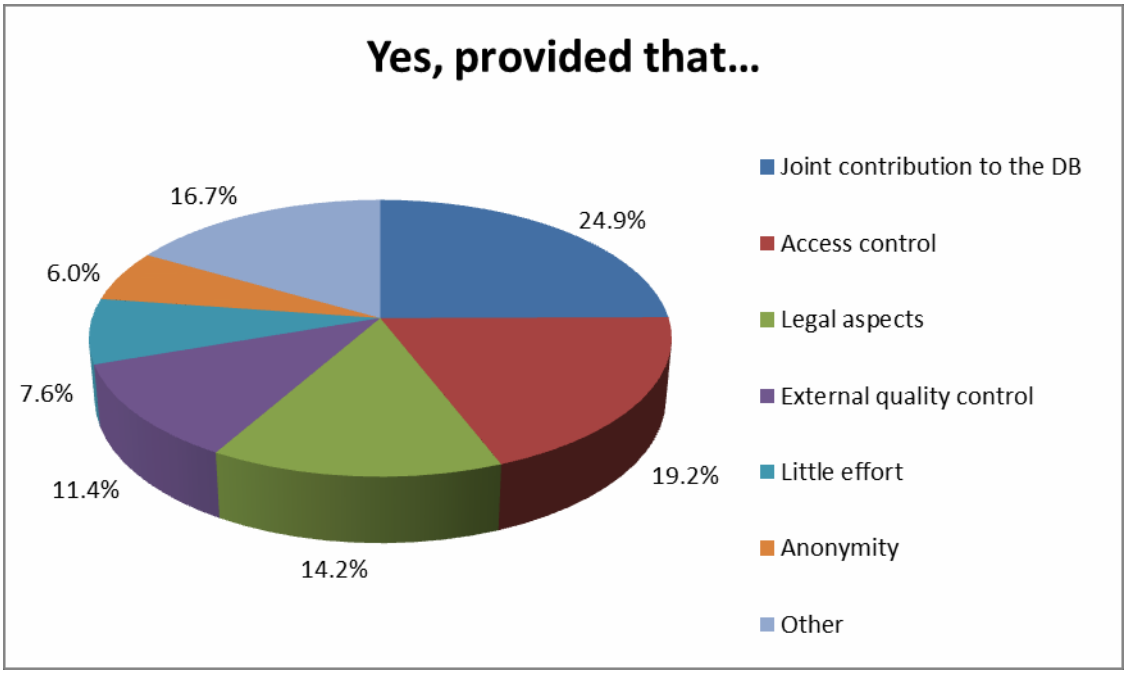

Figure 5. Sharing terminology (Yes, provided that…)

The most frequently named reasons for not being willing to share one's own terminological collections were legal restrictions (48.6%), poor quality of data (22.0%) and the fact that personal terminology collections are one's own assert, i. e., unwillingness to benefit the competitive colleagues (16.5%).

Conclusion

The started work has laid the ground for the fruitful cooperation in the acquisition, processing, sharing and reusing of multilingual terminology resources within academia and business in different usage scenarios and various applications.

Although there exist international standards and well-established best practices for terminology work and terminology management, the practical implementation of both can vary considerably in specific organizational environments and for given professional tasks. The results of the survey provide the basis for the functional and technical specification of the TaaS platform.

The work within the TaaS project has received funding from the European Union under grant agreement n° 296312.

Bibliography

Massion, F. (2007), „Управление терминологией: роскошь или необходимость?" Профессиональный перевод. Выпуск 12, 2007.