Terminology: Don't only collect it, use it!

Vesna Lušicky, University of Vienna Tanja Wissik, University of Vienna

The cost of collecting, elaborating and maintaining terminological resources is enormous in terms of financial, human and time resources. Since terminological work is an on-going process, especially in the legal domain, it is not enough only to collect and store terminology resources. Already existing resources should and could be (re)used and processed to create more complete, higher quality resources. Thus, it is required to add new languages or merge different terminological resources. A semi-automatic method would facilitate the merging of separate resources and increase the reuse of terminological databases.

The LISE (Legal Language Interoperability Services) project, financed by the European Commission in the ICT PSP program, supports processing of already existing terminology resources together on a collaboration platform. The goal is to achieve interoperability across terminology in the same domain and in one or more languages and to facilitate the reuse of terminological resources. In the long-term perspective, this approach also aims at assuring quality of terminological databases.

The focus of the project is twofold: to develop the LISE Service, that is the collaboration platform with the integrated tools, and to investigate terminological workflows. This paper presents the current status of the project.

1. Introduction

Without high-quality and standards-based terminologies, it is impossible to reach precision, efficiency, and transparency within and across any services, processes or systems especially in legal and administrative domain. The cost of collecting, elaborating, maintaining and modifying terminological resources is enormous in terms of financial, human and time resources.

Since terminological work is an on-going process and does not end with the creation of the resource, especially in the legal domain, it is not enough only to store terminological resources. Already existing resources could and should be (re)used and processed to create more complete, higher quality resources. Thus, it is required to add languages to existing terminological resources or to merge different terminological resources. This is a time consuming task and often the reason for keeping the resources separate.

The lack of interoperability between different terminological data is another crucial issue that limits the reuse. A semi-automatic method would facilitate the merging of separate resources and increase the reuse of terminological databases. Terminology management systems are mainly used for data storage, but many other activities, such as collecting information and the majority of the decision-making processes, such as expert consultation, discussion and revision, take place outside the terminology management system and can thus remain undocumented and non-traceable.

2. LISE project

In this paper, we present the LISE project (Legal Language Interoperability Services)[1], financed

[1] For more information on the project please visit the project website www.lise-termservices.eu.

by the European Commission in the ICT PSP programme. It addresses the need for such methods and faces the challenges of interoperability, cleaning, merging and harmonising of already existing terminological resources embedded in the whole terminological workflow. The project is carried out by a consortium of academic, industrial and institutional partners[2].

The focus of the project is twofold: on one hand the project conducts research on workflows and best practices in inter-institutional terminology work with the aim of identifying the routines and procedures in terminology management, specifying the relevant stakeholders and highlighting the needs of the stakeholders involved in order to streamline and improve collaborative terminology work. On the other hand, the project focuses on the development of the LISE Service, i.e. the collaboration platform with the integrated tools. One of the main goals of the LISE project is to work with existing termbases and not to build termbases from scratch.

3. Research on terminology workflows

3.1 Scope and Methodology

The research on terminology workflows was conducted to gain insights into real-life terminology work in the domain of law in different institutions and bodies. The workflow analysis should also provide the basic information for the needs analysis of the different stakeholders involved in the process. It is crucial to analyse and to understand terminological workflows in various institutions and real life scenarios in order to compile the best practices in the field and to fully integrate new tools into the existing workflow.

Two different research methods were chosen for the analysis: an online questionnaire and semi structured expert interviews. In this part we will only give an insight into the expert interviews. Institutions, organisations and bodies that carry out terminology work and/or manage it, were individuated for the interviews in order to obtain a representative collection of the main scenarios in which legal terminology is elaborated and managed. 16 institutions gave their consent to conduct interviews with the correspondent experts (cf. Chiocchetti/Ralli 2012: 10f). In the following when referring to interviews, they are identified by the abbreviation INT and the consecutive number.

The sample consists of international institutions (e.g. FAO), supranational institutions (e.g. EU Institutions and bodies), governmental bodies (e.g. ministries of foreign affairs, Sektion Terminologie der Schweizerische Bundeskanzlei), regional bodies (e.g. Zentraler Terminologiedienst des Staatskantons Bern) and others (e.g. TERMCAT; National Bank).

The interviews were carried out between December 2011 and June 2012. The interviews were recorded, provided that the interviewees granted their consent to do so. Afterwards the interviews were transcribed (cf. Chiocchetti/Ralli 2012: 10f).

3.2 Highlights from the needs analysis

The interviews have shown that the workflows differ from one institution to another (cf. Chiocchetti/Ralli 2012) due to various factors, inter alia: orientation of the terminology work (e.g. standardising or translation-oriented), organisational structure (e.g. single terminologist,

---

[2] University of Vienna, European Academy of Bolzano/Bozen; ESTeam, CrossLang and the Austrian Parlamentary Administration.

team), working profiles (e.g. terminologist, translator/terminologist, lawyer-linguist), number of stakeholders involved (e.g. in-house, intra-institutional, inter-institutional), stages in text/translation production (before, during and after text/translation production), language (monolingual, multilingual) (cf. Wright/Budin 1997: 1f).

The interview partners stated their views on terminology work, their needs and wishes. The need was identified in the cases in which the interview partners have either explicitly or implicitly postulated the gap between „what is" and „what should be" (cf. Witkin et al. 1996). For this study, a needs analysis based on the interviews was conducted. The main needs addressed can be divided into the following abstracted thematic fields and subfields, as also shown in Fig. 1. The outline is followed by a concise assessment of needs.

A Maintaining a termbase

A1 Batch export - review – batch import (INT13)

A2 Discrepancy between the working methods of consolidation and workload (INT6; INT7)

A3 A large number of doublettes (redundant entries), redundant information; missing information (INT7)

A4 Basic collaborative work via e-mails, spread sheets, word processing formats and meetings (INT6; INT7; INT9; INT15)

A5 Manual correction of misspellings or similar mistakes (INT13)

In order to maintain their termbases, a large number of organisations reported using word processing formats and spreadsheets when updating batches instead of working directly in the termbase. Termbase entries are updated and corrected manually one by one or in relatively small batches (cf. INT13; cf. Chiocchetti/Ralli 2012:31). Misspellings are corrected directly, manually in the termbase. One interview partner reported that the current procedure of working on small consolidation projects with current methods does not yield satisfying results in the foreseeable future. Redundant entries and missing information have been reported as a common and time-consuming problem to manage.

B Harmonising and merging termbases

B1 Harmonising (INT2; INT9; INT4)

B2 Merging (INT2; INT6; INT17; INT15; INT10; INT11; INT13; INT12)

The need for more time and focus on harmonising and consolidation was repeatedly stressed, especially by the interview partners that work with large termbases. The need for merging and exchanging of termbases was addressed by some interview partners, but they also stated the issues of copyright (INT5; INT6; INT9; INT13) and of the complexity of the inter-institutional agreements to be the biggest obstacles in their endeavour. Many interview partners expressed their reluctance towards the exchange of data (INT2; INT6; INT8; INT14). The main reasons stated were either that the process is too time-consuming or that that the organisation might run the risk of losing or corrupting its data (INT8).

Data were reported to be merged manually or semi-automatically, and only in few cases fully automatically. One in five organisations stated to have no experience with data merging (cf. Chiocchetti/Ralli 2012: 32). An interchange format is needed in order to merge databases, e.g.

TBX (INT2; INT17) or MARTIF (INT5). About half of the organisations use such standard formats (cf. Chiocchetti/Ralli 2012: 32).

C Enriching a termbase

C1 Limited usability of the monolingual entries in a multilingual termbase (INT6)

C2 Crosscheck between translation memories and termbases (INT1; INT13)

C3 Language specific, subject-oriented projects (INT9; INT17)

More than half of the organisations in the study perform the term extraction manually and a small percentage never carries out any extraction, since they obtain their terminology from other sources (cf. Chiocchetti/Ralli 2012: 30). The question of usability of monolingual entries in a multilingual termbase was raised and there is a clear need for enrichment. Additional languages were reported to be added manually or, in rare cases, by means of automatic processing. Some interview partner reported that they obtain language specific data for enrichment from other institutions or organisations under specific circumstances (INT9; INT17). According to some organisations, they would be interested in conducting a crosscheck between their translation memory and termbases (INT1; INT13).

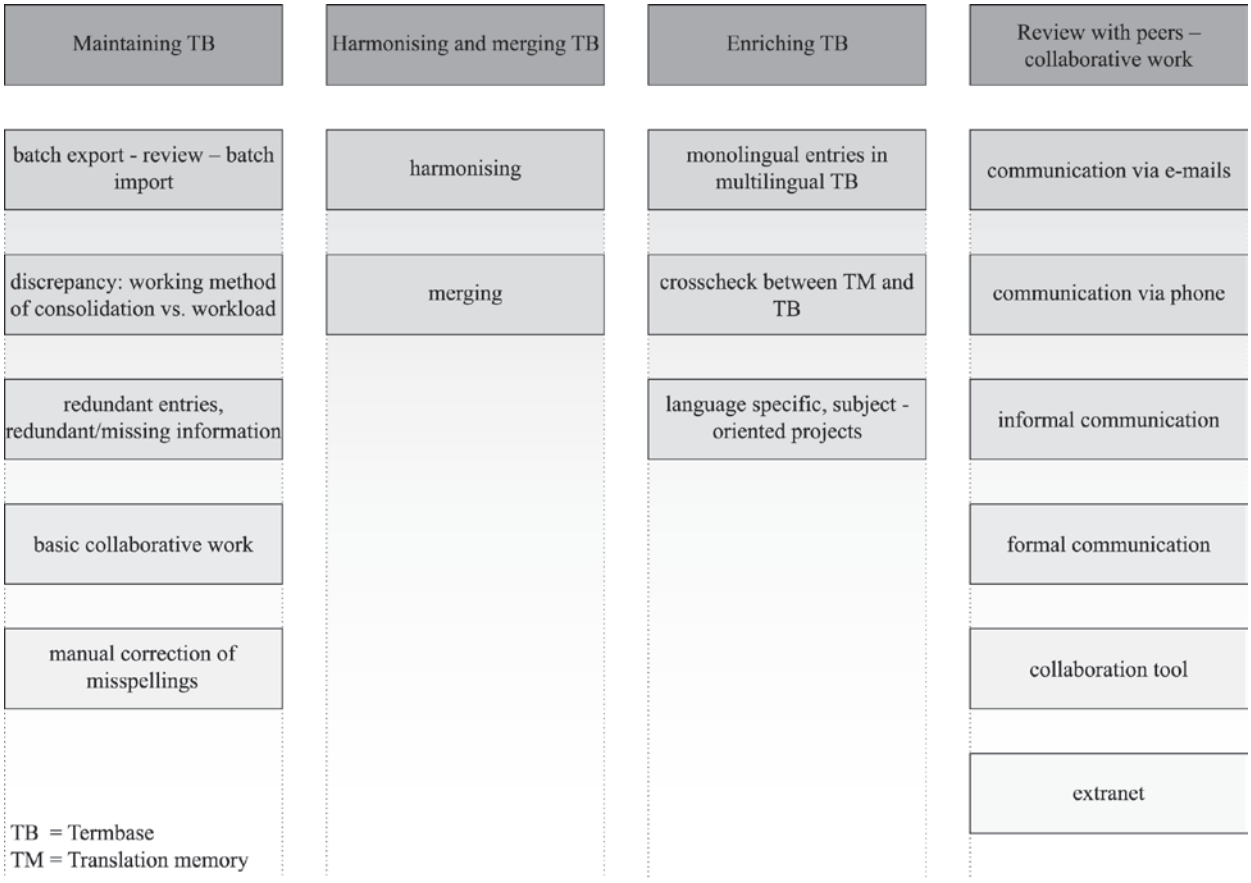

Fig. 1: Thematic fields and subfields identified in needs analysis, based on interviews conducted within the LISE project

D Review with peers – collaborative work

D1 Communication via e-mails (INT6; INT7; INT8; INT9; INT4; INT13; INT15)

D2 Communication via phone (INT2; INT9)

D3 Formal meetings (INT13; INT14; INT15; INT3; INT15)

D4 Informal communication (INT6; INT2; INT4; INT1; INT15)

D5 Collaboration tool- wiki (INT17)

D6 Extranet (INT15)

As stated above, interview partners reported collaborating with their peers on different aspects of termbase maintenance and in reaching out to the domain experts. Although the majority of the interview partners reported that they have limited human resources that work under notable time strain, they rely on various forms of informal means of obtaining information that remains undocumented and limited in dissemination, such as e-mails, informal meeting and phone conversations. Only one interview partner uses an on-line collaboration tool (INT17). Some interview partners suggested that the problem of the limited time and resources could be partly solved by enhancing the access to peers, materials, tools and experts (cf. Chiocchetti/Ralli 2012: 33-34) and by providing a collaborative communication platform (INT7; INT15; INT17).

The qualitative analysis of the interviews shows that there is a need for support in cleaning, harmonising, merging and enriching termbases. There is no demand for wholly automatic processing but for semi-automatic processes that would speed up the above-mentioned tasks and leave the terminologist in full control of their data (e.g. INT7; INT5; INT13). A structured, traceable and interactive communication and decision-making tool is in demand by organisations that carry out inter-institutional terminology work, by organisations with a complex validation hierarchy and/or a large number of stakeholders involved. The LISE Services described in the following section address these issues.

4. LISE Service Version Two

The LISE service supports processing already existing terminology resources together with a collaboration platform. The goal is to achieve interoperability across terminology in the same domain and in one or more languages and to facilitate the reuse of terminological resources. In the long-term perspective this approach also aims at assuring the quality of terminological databases (cf. Wetzel et al. 2012, Chiocchetti/Wissik 2012a; 2012b).

The tools should be fully integrated into the terminological workflow. Therefore, it is crucial to analyse and to understand terminological workflows in various institutions and real life scenarios as was done with the before mentioned interviews and the survey. The tools provided on the platform allow the semi-automatic processing of data when adding languages, cleaning terminological resources from doublettes or harmonising them. Improving the quality of terminological resources is achieved with the help of the tools that facilitate the following tasks:

1) Removing errors and inconsistencies

2) Filling-in of missing languages

3) Identifying and harmonising of conceptually related entries (cf. Wetzel et al. 2012; Chiocchetti/Wissik 2012a; 2012b).

As a user group oriented project, we work closely with our user groups. The main participants of our user group are the EU institutions and agencies that work with IATE, the EU's multilingual termbase. This user group serves as an example for a complex inter-institutional collaboration. As of spring 2012, LISE is piloted with the IATE data. For the purpose of testing the tools, the LISE project received all the IATE data which were analysed in order to plan and develop a prototype of the ESTeam tools adaptation. In the next step, the subset of data that was selected in collaboration between ESTeam, Eurac and University of Vienna, was processed and tested with the ESTeam tools.

In the following section you can see the ESTeam tools working with examples from the IATE data.

Fig. 2 shows the user interface of the Cleanup tool. The example shows a possible spelling error in German "Etäge" - an "r" is missing. This was correctly detected by the tool. The terminologist or the person in charge of the work/task can accept or decline the suggestion of the system.

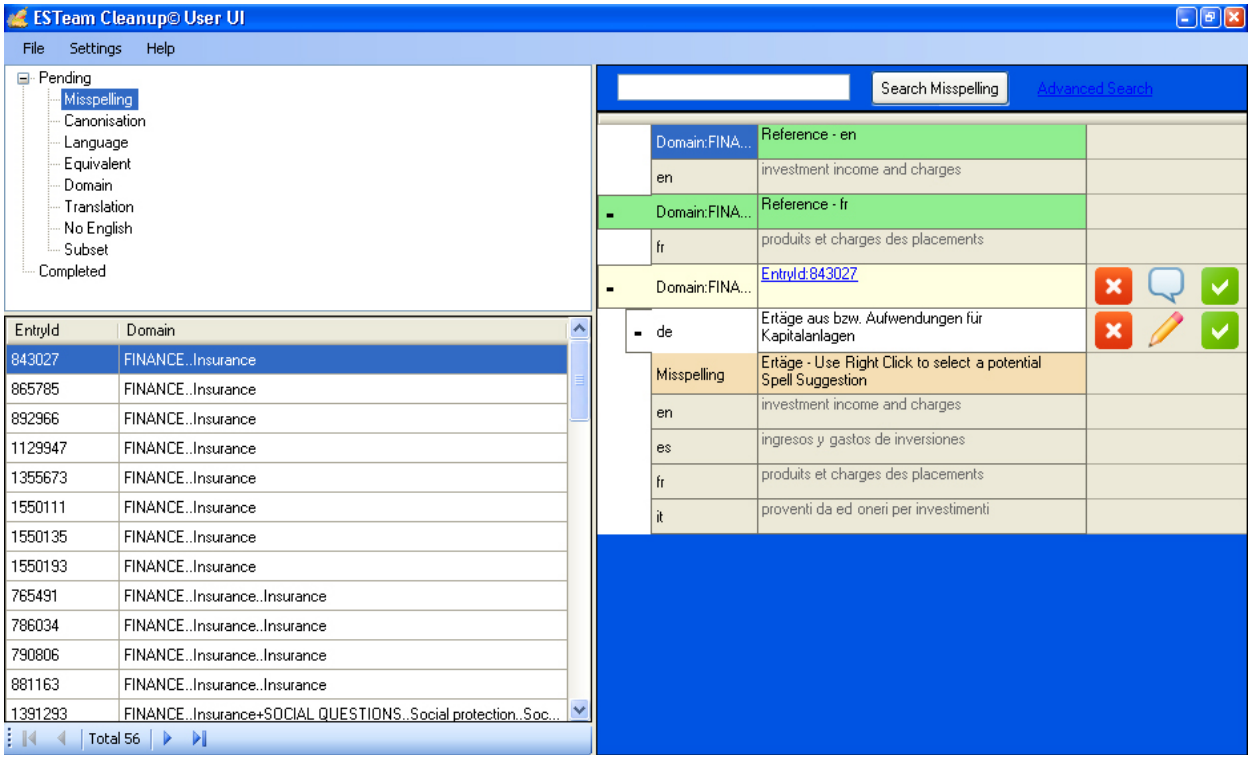

Fig. 2: Cleanup - Detecting spelling errors

Fig. 3 shows the internal harmonisation scenario with the tool OMEO. Three different entries are identified as possible doublettes and are suggested to be unified. As in the previous example, the tool only gives a suggestion, and the terminologist can accept or reject the proposed merging.

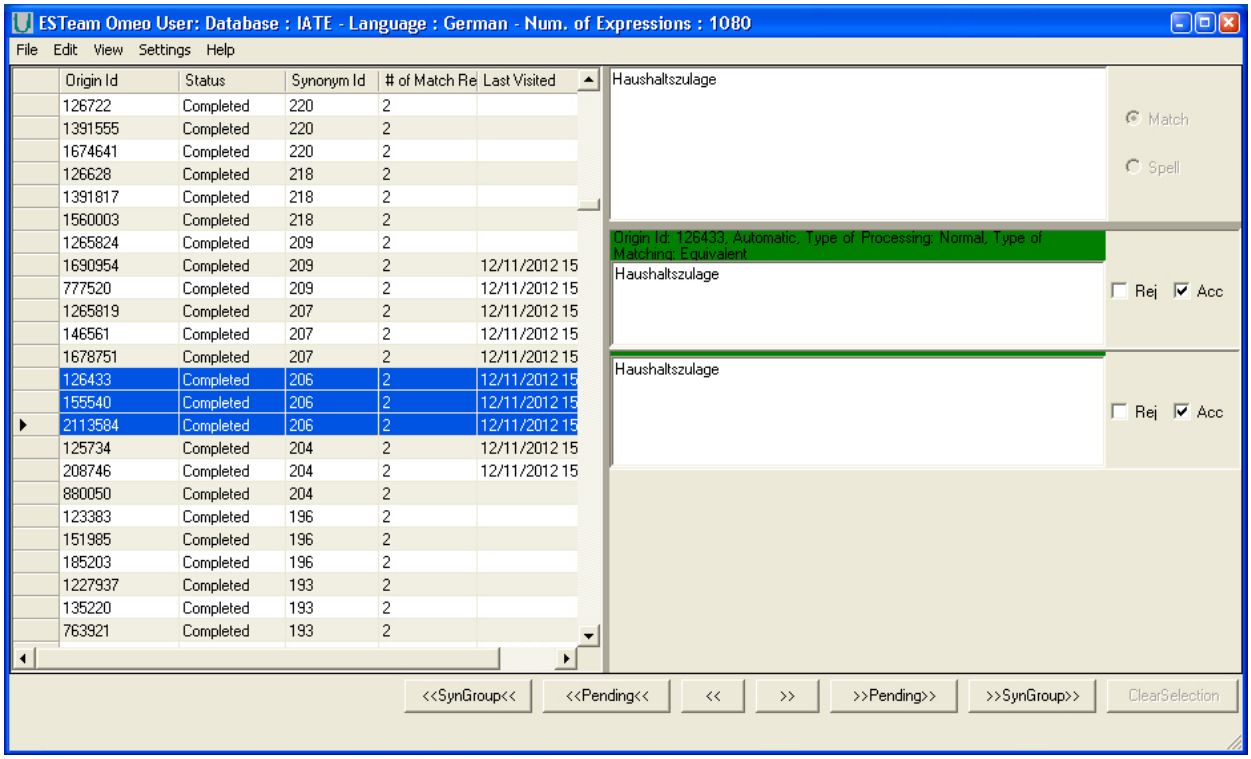

Fig. 3: Internal harmonisation with OMEO

Fig. 4 shows a German translation suggestion from a provided translation memory for an existing entry in English, where German is still missing. The tool used in this example is Fillup.

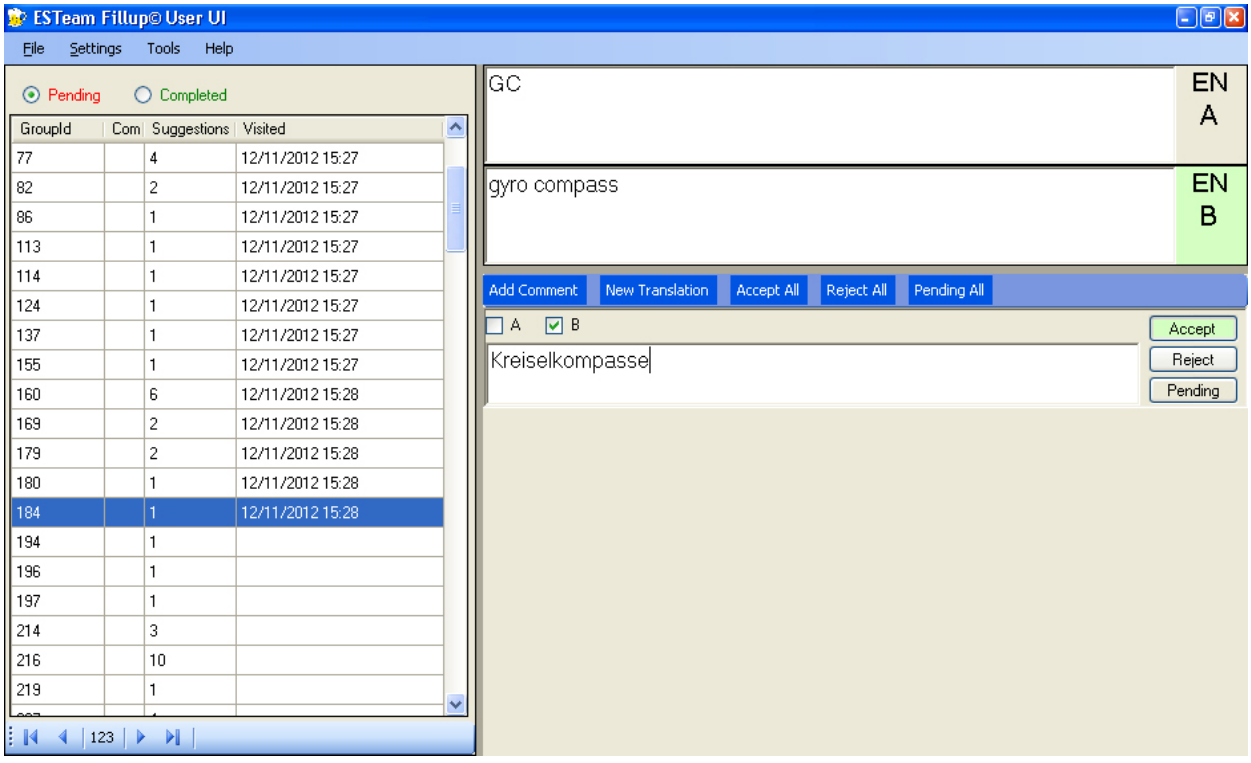

Fig. 4: Adding new languages with Fillup

The above-mentioned tasks require precise communication between all stakeholders involved in the process. Meta-communication, i.e. decision-making on deleting, merging or keeping the entries separate, entering new terms, validating terms and so on, takes place on the collaborative platform.

Within smaller teams the decision making process on terminological issues is usually conducted on the interpersonal level. On the contrary, in bigger teams or even in inter-institutional collaboration groups the communication and the decision-making process are more complex, run on more levels and demand involvement of more stakeholders. During the interviews some interviewees suggested that they would welcome a collaborative tool to communicate and work together with peers and experts and to have access to data, resources and tools all at the same time (see 3.2).

The user-friendly platform within the LISE project would facilitate this meta-communication by providing, inter alia, the following features: creating links to master terminology data, creating topics, adding recipients, defining privacy, adding attachments or voting (cf. Wetzel et al. 2012).

Fig. 5 shows an example of a discussion on spelling variants on the LISE collaboration platform.

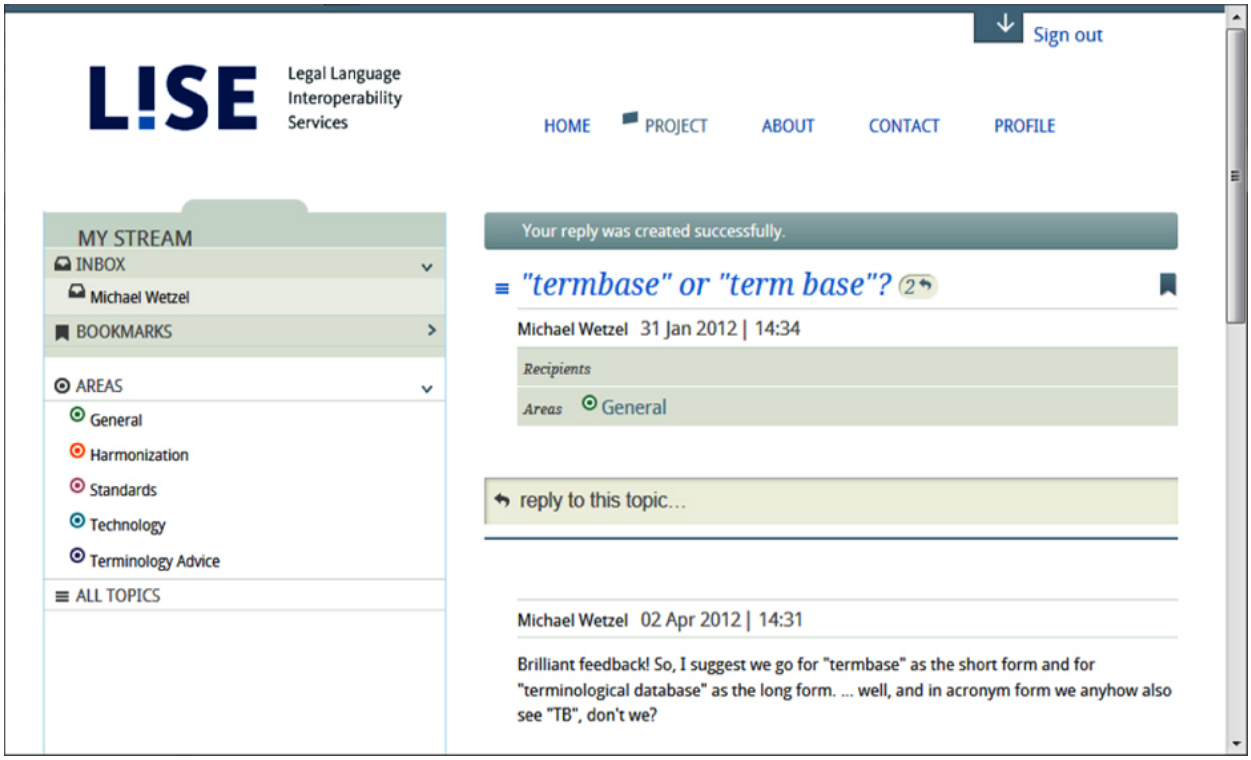

Fig. 5: LISE Collaboration platform

5. Conclusions and outlook

The creation of terminology resources and the development of suitable technology to make use of them is a rather costly and daunting task. Therefore, a user-friendly approach to re-using, cleaning-up, merging, harmonising and enlarging is in high demand. The LISE project is a possible solution for an inter-institutional collaborative approach to terminology processing and

re-using. The prototype of the LISE Service Version Two is available under https://app.lise-termservices.eu/ via registration. By the end of January 2013 the final prototype of the LISE Service Version Three will be available and will enter the evaluation phase with a user group in January. Volunteers from the EU-institutions and agencies involved in terminology work and in the IATE routines will test the LISE Service Version Three.

6. Acknowledgements

The LISE project has received funding from the European Community (ICT-PSP 4th call) under Grant Agreement n° 270917.

7. References

Wetzel, M., Chiocchetti, E., Wissik, T. (2012). Putting Together Apples and Oranges: The LISE Tool Suite for Collaborative Terminology Work. In: Lucas Soares, António & Costa, Rute (Hrsg.) Proceedings ColabTKR 2012 - Terminology and Knowledge Representation Workshop, Prae-Workshop of LREC 2012, 1-6. Available under: http://www.lrecconf.org/proceedings/lrec2012/workshops/08.LREC 2012 Terminology Proceedings.pdf

Chiocchetti, E. & Wissik, T. (2012a). Zusammenführen und Harmonisieren von rechtsterminologischen Datenbeständen: Das LISE (Legal Language Interoperability Services) Projekt stellt sich den Herausforderungen kollaborativer interinstitutioneller Terminologiearbeit. In E. Schweighofer, F. Kummer & W. Hötzendorfer (Eds.), Transformation Juristischer Sprachen. Tagungsband des 15. Internationalen Rechtsinformatik Symposions (IRIS 2012). Salzburg: Österreichische Computer Gesellschaft, 261-268.

Chiocchetti, E. & Wissik, T. (2012b). Harmonising and merging different terminology collections: the LISE term tools. In Aguado de Cea et al. (ed). Proceedings of Terminology and Knowledge Engineering Conference 2012 (TKE 2012). New frontiers in the constructive symbiosis of terminology and knowledge engineering, 19-22 June, Madrid, pp. 310-313.

Chiocchetti, E. & Ralli, N. (2012). D3.2 Report Workflow Adaptation for LISE. Available under: http://ec.europa.eu/information_society/apps/projects/logos//7/270917/080/deliverables/001_LISEdeliverable32.pdf

Witkin, B. R. & Altschuld, J. W. (1995). Planning and Conducting Needs Assessments. Sage Publications: Thousand Oaks, CA.

Wright, S.E. & Budin, G. (1997). Introduction. In S.E. Wright & G. Budin (Eds). Handbook of Terminology Management. Vol. I, Basic Aspects of Terminology Management. Amsterdam, Philadelphia: John Benjamins, 1-12.