



Languages for Special Purposes in a Multilingual, Transcultural World

Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria

<http://lsp2013.univie.ac.at/proceedings>

When IATE met LISE: LISE clean-up and consolidation tools take on the IATE challenge

Paula Zorrilla-Agut

Cite as: Zorrilla-Agut, P. (2014). When IATE met LISE: LISE clean-up and consolidation tools take on the IATE challenge. In G. Budin & V. Lušický (eds.), *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria*. Vienna: University of Vienna, 536-545.

Publication date: July 2014

ISBN: 978-3-200-03674-1

License: This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/4.0/>. This license permits any non-commercial use, distribution and reproduction, provided the original authors and source are credited.



When IATE met LISE: LISE clean-up and consolidation tools take on the IATE challenge

Paula Zorrilla-Agut

*Terminology Team, Translation Centre for the Bodies of the European Union
Luxembourg*

Correspondence to: Paula.Zorrilla_Agut@cdt.europa.eu

Abstract. In 1999, the European Union translation services undertook a rather ambitious endeavour, namely, the “Inter-Active Terminology for Europe” (IATE) project was launched with the objective of creating a single terminology database for all the EU institutions and agencies. The plan was to merge all terminology resources into a web-based and fully interactive database to be shared by all the EU language staff.

In 2004, IATE officially replaced the existing legacy termbases. Although the project has achieved most of its objectives, merging several terminology resources also brought problems, or simply made them more visible. For example, users may find duplicate entries or quality problems due to character conversion issues, missing key data (domain, source), misspellings, etc. Despite the efforts to automate the detection of problematic content, the consolidation of legacy terminology in IATE still remains a challenge for all participating services.

In 2011, the IATE project joined the LISE user group. LISE provides IT tools conceived to tackle the kind of issues IATE is faced with, including semi-automatic consolidation of linguistic resources (detection of duplicates) and identification of other possible quality problems, i.e. misspellings, mistranslations, missing domains, wrong language, etc. In 2013, EU terminologists were able to test these tools and evaluate their potential contribution to the IATE consolidation activities. This paper provides more details on all of the above issues.

Keywords. Automation, clean-up data, conversion errors, data consolidation, domain, doublets, duplicates, EU institutions, IATE, legacy data, legacy terminology databases, LISE project, LISE tools, redundant concepts, terminology database maintenance.

1. The IATE project

1.1. Background

The EU database IATE was created almost 14 years ago with the aim to merge the terminological resources of several EU institutions and bodies into a web-based and fully interactive database to be shared by all the EU’s language staff. Until then, most of the EU institutions and bodies with a translation service had developed and maintained their own terminology resources; however, the data were not fully shared (e.g. TIS by the Council of the European Union, Euterpe by the European Parliament, Eurodicautom by the European Commission, and Euroterms by the Translation Centre for the Bodies of the EU).

The IATE project was coordinated at the beginning by an interinstitutional working group, and since 2009 by an IATE Management Group with representatives from all the project partners: the European Parliament, the Council of the EU, the European Commission, the Court of Justice, the Court of Auditors, the European Economic and Social Committee, the Committee of the Regions, the European Central Bank, the European Investment Bank, and the Translation Centre for the Bodies of the EU.

After defining the technical specifications of the future IATE and reaching a consensus on the data structure (which needed to cater for different needs), working methods and content from the different partners, the development phase started in 2000 and the first legacy data was imported in 2003.

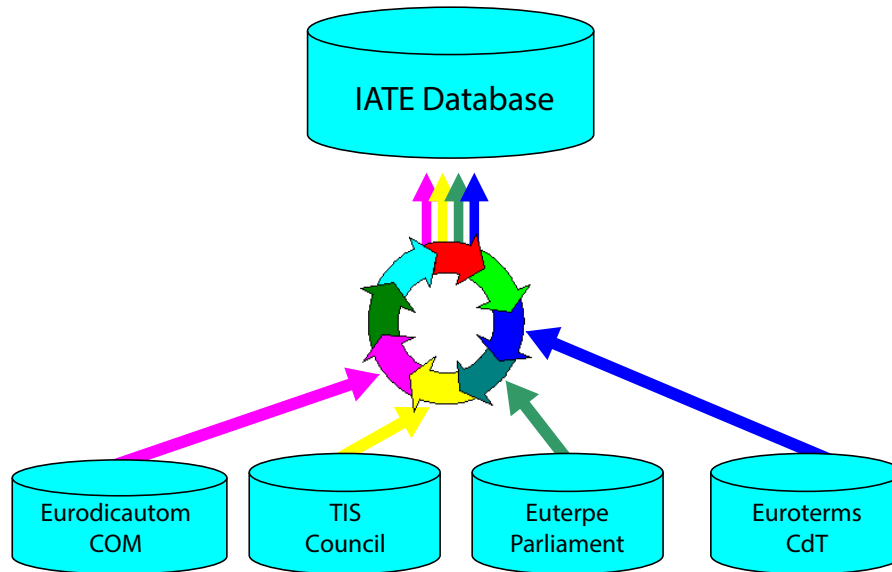


Figure 1: Legacy databases from several EU institutions were merged into IATE

Following a test period with a restricted number of users, in the summer of 2004 IATE officially replaced the existing legacy databases and became the sole and central repository for EU terminological data, thus making relevant terminology available in a shared and interactive way. Cooperation and coordination among the different EU services became essential and reduced the duplication of efforts thus making terminology work more efficient. Several interinstitutional working documents were drafted to agree on a set of best practices and guidelines with a view to a common and harmonised approach to terminology work in IATE.

The closing of the online access to Eurodicautom also increased the demand for an open version of IATE for the general public, which was released in June 2007. The IATE public version contains validated and non-confidential data in the official EU languages which is transferred periodically from the IATE internal database, and comprises only consultation features. The internal version is much more powerful with the features of a fully-fledged terminology management system (i.e. consultation, editing, creation, merging, validation workflow, import, export, user management, feedback management, statistics, etc.).

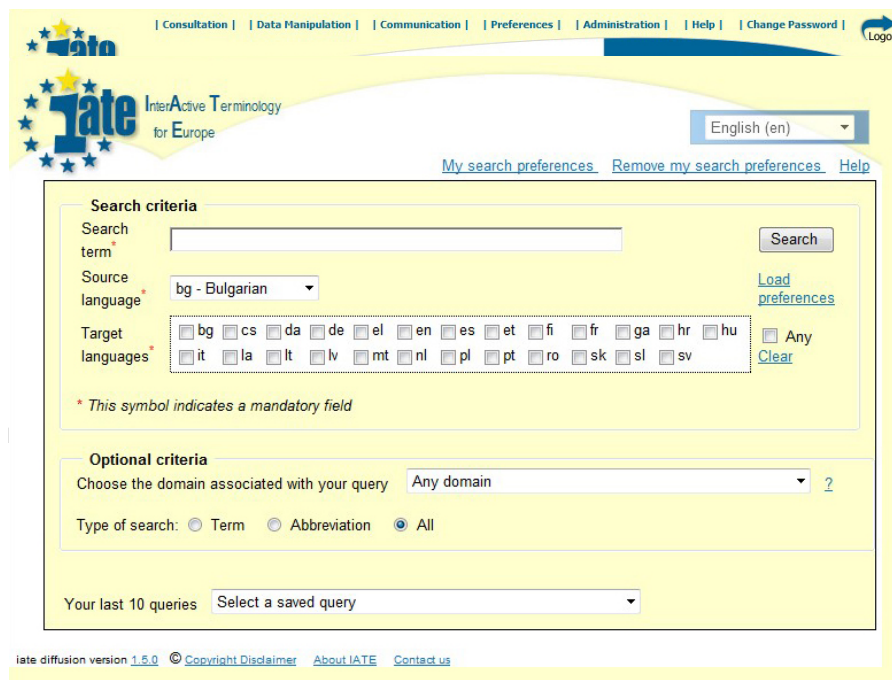


Figure 3: Sample search view in IATE public

1.2. IATE figures

IATE offers at the moment 1.4 million multilingual entries and nearly 9 million terms.

The internal database receives around 40 000 queries per day from EU staff (from mainly translators and terminologists, but also experts and author services), while the public version receives 140,000 queries per day from nearly 200 countries¹.

1.3. Evolution of IATE data

The huge amount of data in IATE calls for ongoing consolidation and updating tasks, which are very challenging given the existing resources. There is a large amount of legacy data that has neither been revised nor updated since it was imported into IATE. There are several reasons for this: domains and needs have evolved and some older collections are no longer relevant (in the domain of IT, for example). The availability of online resources since the advent of the digital era has also changed the approach towards “added value” of data and what should and should not be in IATE. The multilingualism of IATE data –which reflects the multilingual policy of the EU and the different enlargements–, has also expanded significantly, from 11 languages in 2000, when IATE was conceived, to 24 languages in 2013.

Therefore, efforts are not only being put into consolidating duplicates and updating legacy data, but also into covering new domains, dealing with ever-increasing specialised terminology not easily sourced in other authoritative resources, and reducing the imbalance between pre- and post-2004 languages.

1.4. Ownership

Since the conception phase of IATE, data ownership has been a key issue in the management and visualisation of terminology data. The approach to ownership has evolved over the years in line with increased cooperation and common work in the database. IATE was conceived as a common platform but, in the initial phase, data were still grouped by institution (one entry belonging to one specific institution). This meant that any updates, deletions or merging requests were sent by e-mail or via the comments feature in IATE (called ‘marks’) to the owner of the data who would then implement the request. Over time, the consolidation efforts have led to the deletion of many duplicates and the merging of data in a single “primary” entry. The result is that an increasing number of mixed entries can now be found in IATE (an entry consisting of language equivalents provided by different institutions). A major improvement was made early in 2012 with the introduction of the interinstitutional update, which is a more efficient way to improve the quality and completeness of IATE entries. With this feature, any institution can update another institution’s data, and these changes would be submitted for validation by the institution owning the data. This was followed by further features, such as interinstitutional merging whereby any institution can now merge two duplicates independently of the data owner, and more recently by interinstitutional deletion whereby any institution can delete obvious duplicates without any added value from other institutions. This evolution also reflects the greater levels of cooperation and exchange among EU terminologists and translators.

1.5. Roles and rights

Some other relevant factors which may explain the imbalance in the coverage of complementary data in an entry and the lack of uniformity among different entries –particularly for pre-2004 entries– are the different profiles and needs of IATE internal users. Most EU institutions and bodies have three main types of actors in IATE: translators, language terminologists and central terminologists, with different rights in the database, and also different approaches and needs in their day-to-day work. Translators are mostly focused on equivalents, they have limited time for terminology work and their approach is ad-hoc and mainly limited to feeding equivalents into the database as they translate. Language terminologists are responsible for the validation

of data inserted by translators in their institution in their native language. They also carry out terminology work requested from the central terminology services, update references and contexts and give more importance to added value data (definitions, contexts, notes, etc.). The third group consists of central terminologists, who coordinate terminology work for all EU official languages and have a multilingual overview of the entries. They devote more effort to identifying and consolidating existing duplicates at entry level.

IATE at the moment has over 9,000 internal users, of which 5,300 have editing rights (mostly translators), 2,350 have validation rights (language terminologists), and around 130 have administration or advanced rights².

The range of users and their different needs make it very challenging to reach a fully harmonised set of data in the database.

2. Data in IATE

2.1. Main issues of legacy data

The IATE project has achieved most of its objectives, but merging several terminology resources also raised some thorny issues, or simply made them more visible. For example, users may encounter duplicate entries or quality problems due to character conversion issues when legacy data were imported, missing key data (domain, source), misspellings, etc. Despite the efforts to automate the detection of problematic content, the consolidation of legacy terminology in IATE remains a challenge for all participating services.

The main problems which have been detected –particularly in pre-2004 entries– are the following:

- Duplicates (mainly due to legacy data imported into IATE from different databases, but also the result of semi-automatic imports of collections which include concepts already included in IATE)
- Incomplete entries (terms without any complementary information –such as the source, context, or definition–) and entries without any indication of domain or with an incorrect one
- Misspellings (due to manual data input and the absence of spelling-check feature within IATE)
- Broken hyperlinks
- Obsolete data
- Lexical items (LGP) instead of terms (LSP)
- Non canonical forms
- Phrases instead of terms (particular collections, for example international classifications)
- Multiple terms in a term field (wrong data input)

wet lease		Submit Query	Last queries
Your search returned: 4 Hits. 1		Time: 0.36	
31430	Air transport	CdT	1
en - English	wet lease	CdT ****+ @	📄
de - German	Vermieten oder Anmieten mit Besatzung (Wet lease)	CdT ****+ @	
	affrètement/frètement	CdT ****+ @	
fr - French	location avec équipage	CdT ****+ @	
it - Italian	Wet leasing	CdT ****+ @	
nl - Dutch	wet lease	CdT ****	📄
844044	Air transport	Council	2
en - English	wet leasing	Council ****	
	wet lease	Council ****+ @	
de - German	"wet"-Leasing	Council ****+ @	📄
fr - French	affrètement d'aéronef avec équipage	Council ****+ @	📄
it - Italian	noleggio di aereo con equipaggio	Council ****+ @	📄
	wet leasing	Council ****+ @	📄
nl - Dutch	leasing met bemanning	Council ****+ @	
1880743	Air transport	COM	3
en - English	wet lease	COM ****+ @	📄
	wet leasing	COM ****+ @	📄
fr - French	location avec équipage	COM ****+ @	

Figure 4: Duplicate entries for the air transport concept "wet lease"

31430		Air transport
en - English	wet lease	
de - German	Vermieten oder Anmieten mit Besatzung (Wet lease)	
	affrètement/frètement	
fr - French	location avec équipage	
it - Italian	Wet leasing	
nl - Dutch	wet lease	

Figure 5: Multiple terms in a term field (see French)

211932		FINANCE TRANSPORT
en - English	wet charter	
	wet lease	
fr - French	avion affrété avec équipage, carburant, etc	

Figure 6: Descriptions instead of terms (see French)

Domains: 56 - AGRICULTURE, FORESTRY AND FISHERIES		Note:
Printer Friendly Admin Info History Collections		
de en		
en	COM	de COM
Admin Info History		Definition: Obst von Bueschen und StraeueLern, im Gegensatz zu Baumobst Reference: Haenscl-%berkamp Admin Info History
Term Group: 1	COM	Reliability:3
Term: so/t fruit Reference: QP Admin Info History		Term: Beerenobst Reference: Haenscl-%berkamp Admin Info History
		Term Group: 2
		COM
		Reliability:3
		Term: Kleinobst Reference: Haenscl-%berkamp Admin Info History

Figure 7: Corrupted characters

44874		ENVIRONMENT	CdT	1
da - Danish	Kropsdele og organer, herunder blodposer og stabiliseret blod	CdT ****+ @	📄	
de - German	Koerperteile und Organe, einschliesslich Blutbeutel und Blutkonserven	CdT ****+ @	📄	
el - Greek	Μέρη και όργανα του σώματος περιλαμβανομένων οσκών αίματος και διατηρούμενο αίμα	CdT ****+ @	📄	
en - English	body parts and organs including blood bags and blood preserves	CdT ****+ @	📄	
es - Spanish	Restos anatómicos y órganos incluyendo bolsas y bancos de sangre	CdT ****+ @	📄	
fr - French	déchets anatomiques et organes, y compris sacs de sang et réserves de sang	CdT ****+ @	📄	
it - Italian	parti anatomiche ed organi incluse le sacche per il plasma e le sostanze per la conservazione del sangue	CdT ****+ @	📄	
nl - Dutch	lichaamsdelen en organen, inclusief bloedzakjes en geconserveerd bloed	CdT ****+ @	📄	
pt - Portuguese	peças anatómicas e órgãos incluindo sacos de sangue e conservantes de sangue	CdT ****+ @	📄	

Figure 8: Text segments instead of terms (more appropriate for a translation memory)

3. Data maintenance and clean-up

The maintenance and clean-up of IATE data are the main tasks of the IATE partners, explicitly expressed in the framework policies and/or work programmes of most EU institutions and bodies. The EU central terminology services work in close cooperation to reduce the number of duplicate entries and consolidate and increase the quality of IATE data, although each EU terminology service has its own different priorities in terms of the domains to be covered according to its political agenda, which dictates the translation workload, and often approach consolidation in a different way depending on the amount of staff available.

3.1. On-going maintenance tasks

The maintenance and clean-up tasks carried out by the different EU terminology services are usually launched in line with the following initiatives:

- The so called “consolidation projects” are launched by the central terminology service of an EU institution or body in relation to domains of interest. This is mostly manual and intellectual work, consisting of searching for already existing entries covering a particular concept, choosing one of them as the “primary” entry, completing it with as much information as possible and asking other partners who own duplicate entries to merge any relevant information and delete the duplicate entries. Consolidation projects cover a set of primary entries and its duplicates for a particular sub-domain or collection.
- Ad-hoc consolidations: central terminology coordinators are encouraged to consolidate duplicates that are detected while working in IATE, for instance when launching a search. This consists of updating an entry that will be considered the “primary” entry by asking language terminologists to verify and complete their language if needed, and requesting other institutions owning duplicates to delete them after they merge any relevant information into the primary. If time is not sufficient to tackle this ad-hoc consolidation, terminology coordinators will try to signal the issue at least by adding a comment in the obvious duplicate entries. Language terminologists are also encouraged to inform their central terminology service when they come across duplicates so that action can be taken at entry level.
- Through statistics: central terminology services are provided with a quarterly list of the most searched terms both in IATE internal and IATE Public, which allows them to identify potential areas of interest.
- Through basic exports: terminologists can also proceed with batch clean-up and updates by exporting sets of data that comply with certain criteria, such as low reliability, no reference, a specific reference or note that needs to be updated, a missing domain, etc.
- Through advanced exports, carried out by the database administrators.
- For example: potential duplicate entries by language and domain; entries that overlap in several languages; monolingual and bilingual entries which have neither been updated nor completed since a specific date, etc.
- Feedback from internal and external users triggers a correction, update or a consolidation effort of some kind.

3.2. IATE features for clean-up and maintenance tasks

Depending on their role, IATE users have at their disposal a range of features to help them to clean and update existing data.

- Merging: IATE offers terminologists the possibility to merge two entries, which basically consists in copying the selected content of a secondary entry into an entry considered as more complete or reliable into which the content is merged. When merging two entries, the terminologist has the possibility to insert, ignore or concatenate the different fields into the primary entry for each of the existing languages, with the option to preview the result before completing the action. Once the merging is completed, the secondary entry is not automatically deleted, but it is up to the central terminologist to delete it manually.
- Batch update: apart from the individual updates, terminologists with specific rights can run a batch update, which consists of exporting the desired data to Excel format (not all fields are available for a batch update), modifying the data in the Excel file and reimporting the modified data back into IATE by overwriting the previous data. This feature enables global replacements to be made, broken links to be corrected in one go, etc.
- Deletion: currently IATE users can delete data at term, language or entry level manually. Deleted data are sent to a recycle bin and can be restored later if needed. Any batch deletion is done by the IATE database administrators for security reasons.

Apart from these mechanisms and a duplicate detection feature that runs when modifying or inserting a term, IATE does not integrate at the moment any automated mechanism that would allow users to run a more advanced quality verification (spelling issues, potential duplicates per domain, broken links, empty domain or term reference, etc.).

3.3. Adaptations in IATE

Since 2004, IATE has evolved in order to cater for the maintenance and consolidation needs of its users following interinstitutional discussions and suggestions from terminologists. Some major developments include the following:

- Batch import feature, which allows terminologists to carry out batch updates directly from the IATE interface.
- Creation of PreIATE, which is a repository within IATE, which contains raw material. PreIATE gives IATE partners more flexibility to import certain collections and encourage terminologists to insert “work in progress” data. Terms marked as PreIATE are not transferred to IATE Public.



3549812	
en - English	persons of concern <i>pre</i> 
it - Italian	persone in stato di bisogno <i>pre</i> 

Figure 9: Entry with two PreIATE terms

- Primary entries: consolidated entries, which have been thoroughly revised and updated in all languages, are marked with a star displayed at the entry level. Any duplicate entries should be merged into the primary and deleted. The “primary” mark is also used when sorting the results, giving priority to those entries in the hit list.

3518459	FINANCE	★ Council	16
bg - Bulgarian	изходна стратегия за финансовия сектор	Council ★★★★★ @	
cs - Czech	ústup od angažovanosti státu ve finančním sektoru	Council ★★★★★ @	
da - Danish	finansiel exit	Council ★★★★★ @	
de - German	Ausstieg aus der Finanzmarktstützung	Council ★★★★★ @	
el - Greek	έξοδος από τη χρηματοοικονομική κρίση	Council ★★★★★ @	
en - English	financial exit	Council ★★★★★ @	
es - Spanish	salida de la crisis financiera	Council ★★★★★ @	
et - Estonian	finantstoetuskeemidest loobumine	Council ★★★★★ @	
	finantstoetuskeemidest väljumine	Council ★★★★★ @	
fi - Finnish	rahoitustukitoimien purkaminen	Council ★★★★★ @	
fr - French	sortie des programmes d'aide au secteur financier	Council ★★★★★ @	
ga - Irish	scor ar bhearta tacaíochta na heamála airgeadais	Council ★★★★★ @	
hu - Hungarian	a pénzügyi szektort érintő válságkezelő intézkedések leépítése	Council ★★★★★ @	
it - Italian	uscita dalle misure di sostegno pubblico al settore finanziario	Council ★★★★★ @	
lt - Lithuanian	finansinis pasitraukimas	Council ★★★★★ @	
lv - Latvian	finansiālā atbalsta pakāpeniska samazināšana	Council ★★★★★ @	
mt - Maltese	hruġ finanzjarju	Council ★★★★★ @	
	hruġ mill-programmi ta' għajjuna lis-settur finanzjarju	Council ★★★★★ @	
nl - Dutch	financiële exit	Council ★★★★★ @	
pl - Polish	finansowa strategia wyjścia	Council ★★★★★ @	
pt - Portuguese	estratégia de saída no domínio financeiro	Council ★★★★★ @	
ro - Romanian	strategii financiare de ieşire	Council ★★★★★ @	
sk - Slovak	ukončenie finančnej angažovanosti	Council ★★★★★ @	
sl - Slovenian	ukinitvev programov finančne pomoči	Council ★★★★★ @	
sv - Swedish	finanspolitisk exit	Council ★★★★★ @	

Figure 10: Consolidated entry with a primary entry icon

IATE macro, which allows a query to be launched and enables IATE to be fed from Microsoft Word so as to integrate terminology work into the translator's working environment to ensure that IATE is consulted, remains relevant and is constantly updated.

Figure 11: Pop-up window of the feeding feature of the macro

4. IATE meets LISE

The LISE (Legal Language Interoperability Services) project is aimed at enabling data owners in public administrations and translation departments to manage their terminological data on the basis of best practices in interinstitutional, interdisciplinary and multilingual terminology management workflows and using web services to support this work. The tools developed in

the framework of this project can be used for the semi-automatic consolidation of linguistic resources and enable the detection of potential quality problems, i.e. misspellings, mistranslations, missing domains, wrong language, etc., which are the kind of problems typically faced by any big and multi-partner terminology database such as IATE. In 2012, the LISE project sent a request to the EU institutions to include IATE in the project's user group and to use IATE terminology to demonstrate the usefulness of the project tools. This request was approved by the Interinstitutional Coordinating Committee for Translations and the project has been followed by the IATE Management Group.

4.1. LISE tests

In February and April 2013, EU terminologists had the chance to test these tools with a set of 67 000 entries covering the domains of social security, social questions, working conditions and insurance. 10 EU terminologists evaluated the three modules as described below for the following languages: DE, EN, ES, FR, PT and SV (not all post-2004 EU languages were supported for the linguistic checks).

The tools to be tested were provided through a desktop application, where testers defined their working languages and a set of IATE entries was pre-loaded (the transfer of data from IATE to LISE was done by the LISE group and not covered in these tests). Below is a more detailed description of the tests carried out with the different modules.

4.2. Clean-up module, addressing the following linguistic and redundancy issues

- detection of spelling mistakes
- canonisation
- language recognition
- mistranslations
- data management errors (missing data)
- additional domains suggested
- overlapping entries (same concept in singular and plural)

The clean-up module was evaluated as being quite reliable for the languages tested and the IATE user group concluded that most of the features offered are useful for a formal, linguistic clean-up of data. One aspect raised by the users was the importance of visualising more data categories (definition, domain, reference, update date) for decision-taking in redundancy-related issues. Users could only evaluate the tools with data from very specific subdomains, and it was not possible to assess whether the redundancy-related features would also present similar results on other wider domains.

4.3. Omeo module, addressing conceptualisation issues

- detection of potential duplicates and related entries comparing anchor language, target languages and regrouping them (from monolingual to multilingual grouping), so that they are later processed together in order to optimise the consolidation work

The tests carried out with Omeo proved that thorough manual (merging) and intellectual work (consolidation of duplicates) was still required; however the fact that Omeo detects not only duplicates but also closely related entries was seen as useful in order to speed up consolidation work. Again, users highlighted the importance of visualising key metadata (owner, primary), of added value fields (definition, context, note) and full multilingual entries (not only two languages) for decision-taking when merging entries (deciding which entry should be considered as primary).

4.4. Fill-up module, helping to enhance and complete terminology entries

- use of TMs to find equivalents for existing terms into languages that are not covered

The Fill-up module, if adapted to several IATE best practices (retrieving also the term source, extraction of contexts, etc.) and used with highly reliable translation memories which contain final versions, could help populate less-represented languages in IATE. Those automatically extracted equivalents could be marked as PreIATE data for further evaluation and validation.

A collaborative platform with discussion, rating and information exchange features was also presented but not tested.

5. Conclusions

The tests allowed the IATE user group to confirm the obvious issues which affect the reliability and quality of part of the data in IATE and the usefulness of semi-automatic tools to speed up the identification and correction of issues.

During the workshop, the IATE user group could make corrections and modify the pre-loaded data, although there was no connection and transfer of updates between IATE and the LISE tools at this stage. Therefore the modules were used as a standalone reporting tool. The interface was rated as simple and user-friendly.

Following these tests, the IATE user group concluded that a quality assurance module which would enable linguistic and conceptualisation issues to be tackled should be ideally integrated into IATE in order to allow a direct editing of the reported issues and avoid data transfer and conversions. Given the multilingual nature of IATE, the linguistic quality assurance module would need to cover the 24 EU official languages.

The tests were seen as a good exercise for raising awareness on certain quality issues in IATE that could be addressed with better level of automation with respect to clean-up and maintenance tasks.

6. Acknowledgements

I would like to thank the Terminology Team at the Translation Centre, the IATE Management Group (and particularly its chair Mr Dieter Rummel) and the IATE Support and Development Team for all I have learned from them in the past years.

7. Notes

¹ Figures obtained from IATE Central Statistics for the second quarter 2013.

² Figures provided by the IATE database administrators in July 2013.

8. References

IATE public version, www.iate.europa.eu, accessed on 15.10.2013.

IATE public brochure, <http://iate.europa.eu/iatediff/brochure/index.html>, accessed on 15.10.2013.

LISE project, <http://www.lise-termservices.eu/>, accessed on 15.10.2013.