[*Translating and the Computer 23*: papers from the Aslib conference held on 29 & 30 November 2001]

The IATE Project - Towards a Single Terminology Database For the European Union

D. Rummel[1] & S. Ball[2]

1. Translation Centre for Bodies of the EU, 1, r. du Fort Thuengen,
L-1499 Luxembourg (Dieter.Rummel@cdt.eu.int)

2. European Parliament Translation Service, L-2929 Luxembourg
(sball@europarl.eu.int)

ABSTRACT

The IATE project was launched in early 2000 for the creation of a single central terminology database for all the institutions, agencies and other bodies of the European Union. By mid-2001, it had reached the prototype phase. It is evident that the attempt of uniting the terminology that has been created in different institutions, with different approaches to terminology and different working cultures, was not an easy task. Although the implementation of the system has, from a technical point of view, already reached a rather advanced stage it is predictable that user feedback during the prototype and pilot phases will still lead to a number of changes. The biggest challenge of the project, however, lies in its introduction in the terminology and translation workflow of the participating bodies. This is illustrated in the second part of the paper by the example of the European Parliament's Translation Service.

PART ONE: CURRENT STATUS OF THE IATE PROJECT

INTRODUCTION

Since 1995 high level representatives of the translation services of the European Union's Institutions and agencies have met in the Interinstitutional Committee for Translation. This committee was set up to formalise contacts and cooperation between these partners that had already existed on an informal level previously. The mandate that governs its activities clearly stresses two aspects: that the partners involved share to a large extent the same problems and face similar challenges; they should thus thrive to find common solutions; whilst underlining, on the other hand, that one translation service is not like another; that due to the role each of the bodies of the Union plays in the life of the Community the practicalities of translation work may differ strongly from one service to another. Still, one of the fields where close cooperation was not only considered in the interest of complying with ever tighter budget lines, but also as offering substantial advantages for the linguistic staff, was terminology.

The availability of terminological resources in the translation services of the Union as such was (and is) far from what one would call unsatisfactory. The "big three", Commission, European Parliament and Council, have each built up powerful terminology databases: the Commission's Eurodicautom, as the oldest and biggest of the institutional databases contains about 1.4 million multilingual concepts. It offers, as do the Council's TIS and the EP's EUTERPE, web-based search interfaces and thus gives access to a vast store of linguistic information to a wide public inside and outside the institutions. The picture looks less bright for the smaller institutions and agencies who in some cases use internal databases (usually in MultiTerm 95) or make do with glossaries in word processor formats. Cooperation on terminological questions and the sharing of information is far from evident even for bodies that need to work closely together like, e.g. the decentralised agencies and the Translation Centre.

There are yet more drawbacks to this situation. The absence of a single point of access to all terminological data makes the lives of translators and other people searching for terminological information difficult. In order to access all available information from the big three databases you would have to learn and use three different interfaces. Attempts to import data from TIS and Euterpe into Eurodicautom to overcome this difficulty have given only unsatisfactory results. Not only do the technical difficulties of the process make regular updates impossible. The difference in the data structures, expression of different terminological cultures and working methods, also lead to a loss of data in the import process.

This fact points to another, more general problem that goes beyond pure convenience for the end-user. The existence of parallel, independent approaches for the creation and maintenance of terminology have made cooperation between institutions and agencies difficult if not impossible. There is no easy way of standardising the usage of terminology between institutions. Problems of inconsistency, redundancy in the data and duplication of work result from this "balkanisation" of the terminology in the European Union.

A study carried out by the IT consultancy company ATOS in 1998 clearly analysed the shortcomings of this situation and concluded that the best remedy was the creation of a single interinstitutional terminology database. After the definition of a common data format all data collected by the different institutions should be merged into this database. But the recommendations of the report went yet further: they stressed the need for wider interinstitutional cooperation in the field of terminology, the reorganisation of terminology activity, reinforcement of staffing where necessary and the build-up of an infrastructure that would allow for cooperative data management.

Acting on the recommendations of the ATOS study, the Translation Centre launched the "IATE" ("Inter-Agency Terminology Exchange") project in 1999; its initial objective was to create an infrastructure for the management of terminology for the Centre and the decentralised agencies of the Union. The other European Institutions later joined this initiative and gave the project its truly interinstitutional status.

The implementation of the IATE project started in January 2000. A consortium of the Greek IT company *Quality&Reliability* and the Danish government research institute *Center for Sprogteknologi* (CST) developed – together with institutional participants – the technical and functional specifications of the European Union's terminology database. In summer 2001 the tests of the prototype of this system were performed. Concepts that have been developed by the participants of various work groups in the phase of system analysis and design have become usable features of the prototype: interactive on-line data entry, a flexible validation system, tools for monitoring, reporting and auditing, advanced user management and modules for large scale data management are operational. However, when we speak of a *prototype* we should be aware that there is still some way to go until this database will be accessible for institutional users and a wider public. The first version of the EU term base made it possible for a group of test users to carry out functional tests, i.e. to check whether the underlining concepts have been implemented correctly and whether they were correct in the first place. A number of aspects of the system, especially the design of the user interfaces, will be subject to considerable modifications in the near future. The screen shots reproduced in this paper are taken from the prototype and should thus be seen as what they are: a glimpse of work in progress and not as a final product. Another two pilot test phases, scheduled for the first two quarters of 2002, will reflect the experience gathered during the prototype test phase and brings us much closer to a system that hopefully combines functionality and user-friendliness.

It is beyond the scope of this paper to give a detailed account of all the different features and modules that have been implemented so far. I will concentrated on following three major aspects of this development:

- **One common database** for all institutions and agencies containing all legacy data;
- **Interactivity,** i.e. the possibility for user to carry out modifications, to add entries directly on the central database and to allow thus their colleague to profit from this work immediately;
- **In-build validation procedures** to ensure quality.

Other features, that can be discussed only briefly, include:

- **management tools,** e.g. for user and data administration;
- **reporting tools;**
- **messaging systems** as communication mechanisms between the actors in the terminology workflow.

LEGACY DATA

Merging the terminology of the existing institutional databases into one single database was a major challenge of the first phase of the project. So far the following databases have been imported into the EU term base: Eurodicautom (Commission), TIS (Council), Euterpe (EP), Euroterms (Translation Centre) and CDCTERM (Court of Auditors). Data from the Court of Justice and the European Investment Bank will be added during a second phase of data loading scheduled for the beginning of 2002. The resources of other European bodies can be added at a later stage as the need arises.

The first achievement of the IATE project is that the legacy data has been physically merged into one relational database[1] without serious loss or corruption of data.

Concepts	1,433,252
Language specific entries	7,793,060
Definitions	815,560
Explanations	51,320
Breakdown by entry type	
Terms	7,111,480
Abbreviations	493,603
Phrases	187,872

Table 1: The prototype of the EU **term base in figures**

This task was challenging not only because of the tremendous amount of data that had to be treated (Eurodicautom alone contains about 1.2 million multilingual concepts); a bigger problem than the actual number of entries was the content of the different databases and the ways in which it is structured: different philosophies of terminology and different historical backgrounds that are expressed in the data stored had to be reconciled. This process involved, in a first step, the definition of mapping rules between the data structures of the existing databases and the new format of the interinstitutional database. This data structure takes into consideration the evolving standards in the field (SALT/MARTIF, GENETER). It adopted a concept-oriented approach; the mono- and multilingual information on each aspect of a concept can be expressed on four inter-related levels of the data structure of the terminological entries:

[1] The technical implementation of the EU term base is based on Oracle 8I RDBMS using Oracle Intermedia for the indexing. The data is stored in Unicode (UTF8).

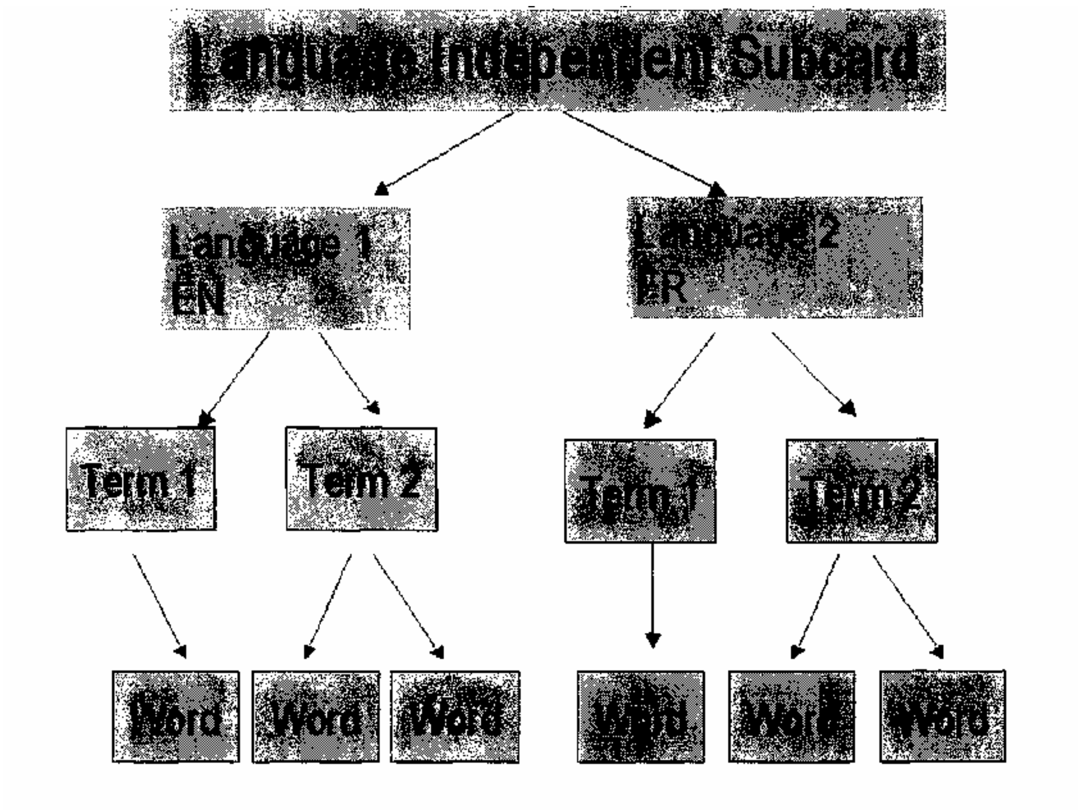


Figure 1: Basic data structure of the EU **term base**

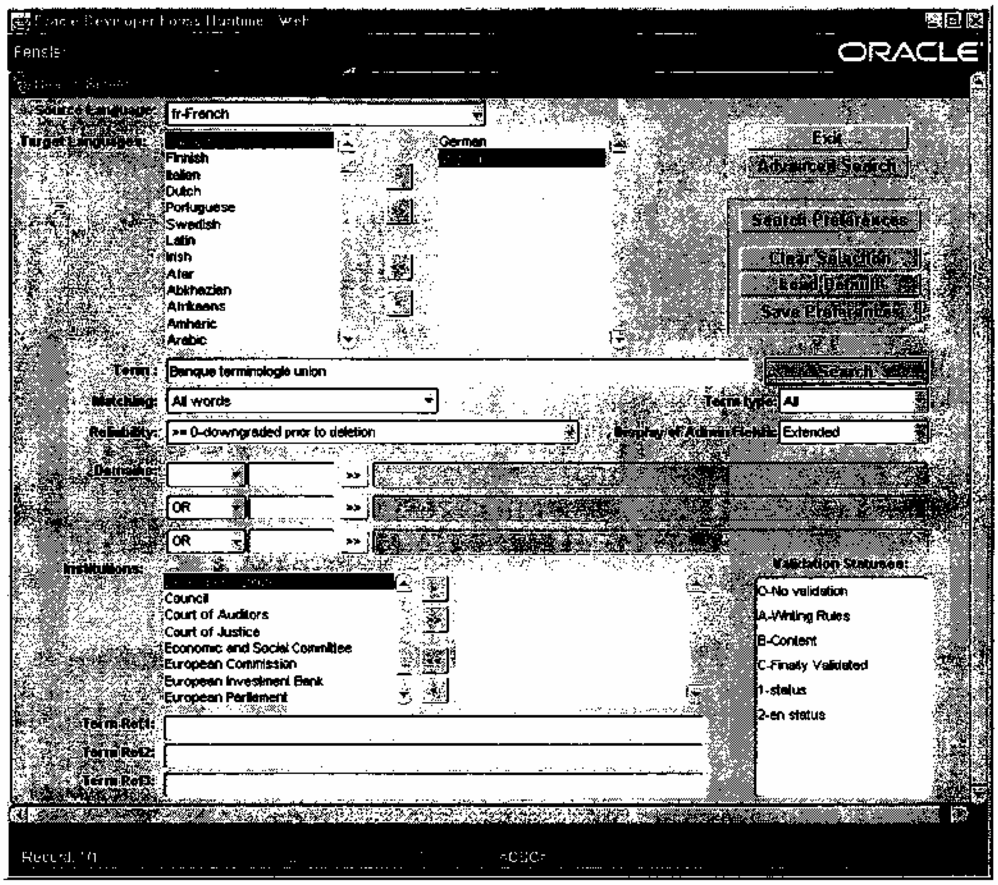
1. the language-independent level can contain all information that relates to the entire concept. "Domain" is the classic example of that type of information. But the database also makes it possible to be more exhaustive: the user can add a domain note in cases when the classification system for domains does not contain a suitable descriptor; collection, problem language, cross references to other entries, origin of the concept and - as we are living in an age of multimedia - links to images complete the language independent level.
2. Beneath this top level, information like definition, explanation and comments can be stored in and for each of the languages the entry contains. This level is enriched by the possibility to add notes on several fields, references to source documents and, again, multimedia files.
3. Each language level may refer to several terms - synonyms of the same concept or abbreviations. A large variety of information can be associated with each of the terms: term type, reference, regional usage, context, customers, links to homonyms etc.
4. Finally the system includes the option to add linguistic information, like part of speech or gender, for each term or each of the words constituting a term.

The first step in the conversion of the legacy data was, as mentioned above, the mapping of all data fields in the existing term bases to a corresponding field in the [ATE structure. It is evident that this was not always a straightforward process. In some cases the idiosyncratic use of certain data fields in the legacy databases made it necessary to apply complex algorithms to come to a satisfactory result. The participants in the project are aware that this first import of data, although successful to a large degree, will have to be repeated as remaining difficulties become apparent. A specific part of the project deals with the problems related to the consolidation and standardisation of legacy data. The following aspects have been detected so far as being crucial for the quality of the legacy data:

- detecting and dealing with perfect duplicates;
- detecting and dealing with partial duplicates (e.g. same subject domain and same term, abbreviation or expression in at least one language);
- identifying and dealing with data of poor quality (e.g. with no definition);
- harmonising/normalising references;
- harmonising the use of standard values for certain fields.

BASIC AND ADVANCED SEARCHING IN THE EU TERM BASE

Creating a "single point of access" to all terminology resources of the Union – one of the major objectives of the LATE project – means providing a user friendly search interface to the central database. During the design phase of the project it became clear that this is not as straightforward a task as it may seem. Not only is "user-friendliness" a term that allows for very different interpretations; but the EU term base must also cater for the needs of very different user groups: translators, terminologists and the European citizen. It also has to take into consideration that for a database of this size very basic search criteria may quickly prove insufficient.



Source Language: fr-French

Target Languages:
Finnish
Italian
Dutch
Portuguese
Swedish
Latin
Irish
Afar
Abkhazien
Afrikaans
Amharic
Arabic

German

Term : Banque terminologie union

Matching: All words

Term type: All

Reliability: >= 0-downgraded prior to deletion

Display of Admin Fields: Extended

Domains:
OR
OR

Institutions:
Council
Court of Auditors
Court of Justice
Economic and Social Committee
European Commission
European Investment Bank
European Parliament

Validation Statuses:
0-No validation
A-Writing Rules
B-Content
C-Finaly Validated
1-status
2-en status

Term Ref1:
Term Ref2:
Term Ref3:

Figure 2: Basic search screen

Minimal search criteria are the language of the search term and the term itself. The user can also specify the target language he or she is interested in. The order in which these languages are selected is repeated in the display of the query results. Other search criteria can and should be added to refine the search results:

Domain classification: The IATE work group responsible for questions related to the content of the database (Data Content Group) decided to adopt the EuroVoc thesaurus for the domain classification of entries in the EU term base. The alternative proposition, the Lenoch classification that is used in the Eurodicautom database, was regarded as a complex, very rich, fine-grained system, that allows for a very precise classification of concepts. This positive characterisation is at the same time the reason why, after some discussions, it was decided to vote for EuroVoc: Lenoch demands expertise in classification. Translators would be able to enter first-level codes, but the allocation of lower-level codes would have to be done by experts. EuroVoc was regarded as offering several other advantages: it exists in all official languages of the Union, includes a list of keywords in natural language and benefits from the support of an interinstitutional mechanism for maintaining and enhancing content. In addition, it is based on the corpus of texts that are created by the Union, i.e. it is centred on our fields of interest.

Matching: Different match operators allow to specify the degree of correspondence requested between search term and matching database entry, e.g. "Containing any of the search terms", "Containing all of the search terms (independently of order)", "exact match", "fuzzy match" and "partial match".

Entry Type: Each entry in IATE belongs to a specific category, e.g. "Term only", "Phrase", "Abbreviation", and "Formula".

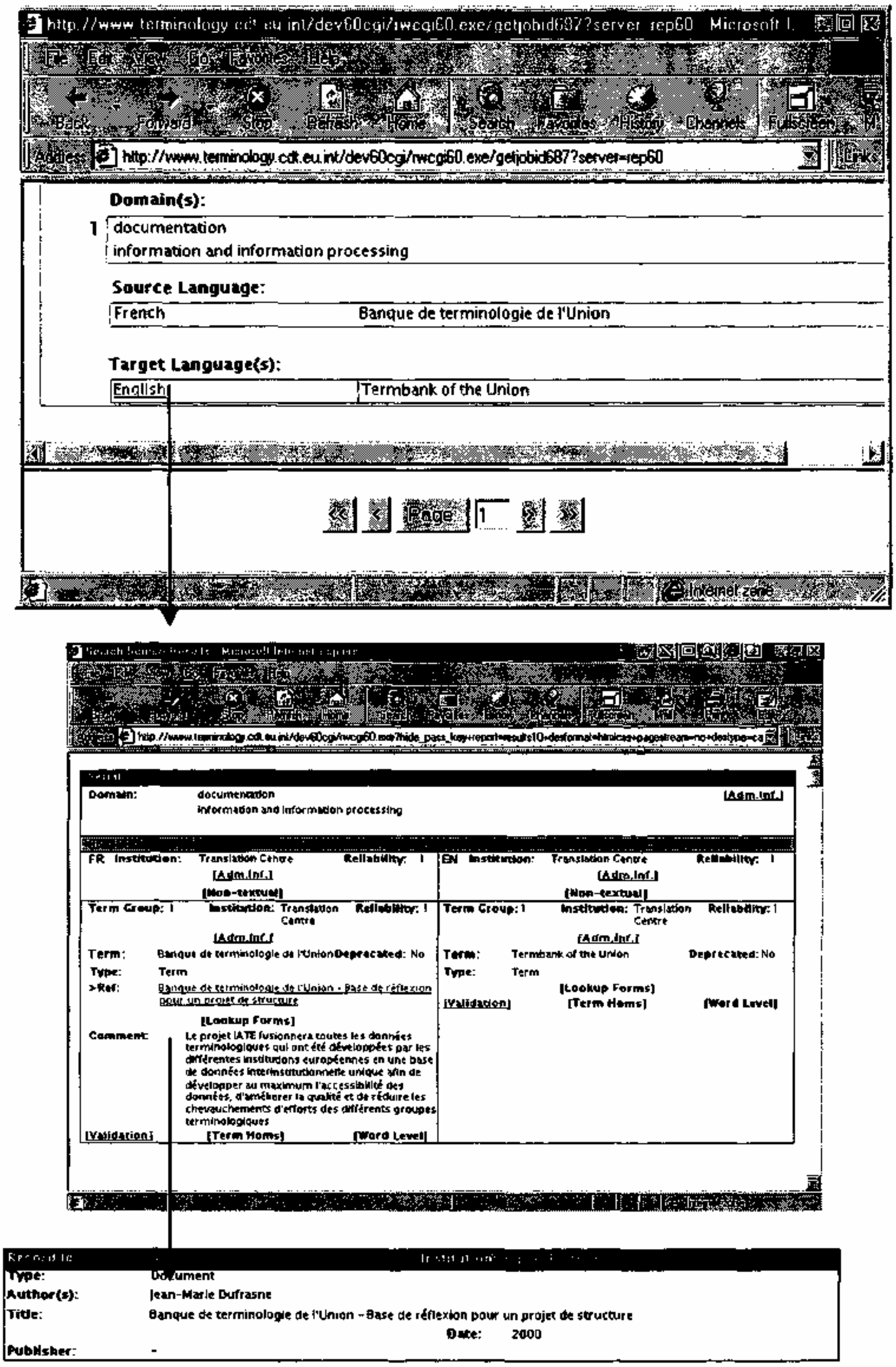
Institution: The idea of "ownership" of data, that might be seen as being a contradiction to the basic idea of one common database, is maintained in EU term base. This is true both for legacy data and for newly created entries. "Institution" as a search and sorting criteria allows translators to focus on the terminology that is used and confirmed in his or her institution if necessary.

Other criteria that allow for fine-tuning the search: Reliability, Validation status.

As any of these criteria may, in any combination, be used by translators for different tasks - different document types, customers or subjects - the system provides for a simple possibility of saving query settings in named profiles. This makes it possible to quickly restore or switch between even very complex search criteria.

The result of a query is a hit list containing some basic information on the concept retrieved: domain, languages, the matching term and its translations. Hyperlinks give access to more specific levels of information. A detailed result display shows all fields of an entry that contain linguistic data. From this screen it is possible to access even more fine-grained elements of the entry.

Besides this basic search facility the EU term base also has to satisfy the need of expert terminologist. An "Advanced search" screen can for example be used to search for entries that contain a specific term *and* a specific translation in one or two other search languages. Finally the system allows for the formulation of search requests in structured query language (SQL).



http://www.terminology.cdt.eu.int/dev60cgi/rwcgi60.exe/getjobid687?server=rep60

Domain(s):

1 documentation
information and information processing

Source Language:

French	Banque de terminologie de l'Union

Target Language(s):

English	Termbank of the Union

http://www.terminology.cdt.eu.int/dev60cgi/rwcgi60.exe?hide_pass_key+report=results10+desformat=htmlcss+pagestream=no+destype=ca

Domain: documentation
information and information processing [Adm.Inf.]

FR Institution: Translation Centre Reliability: 1 [Adm.Inf.] [Non-textual]	EN Institution: Translation Centre Reliability: 1 [Adm.Inf.] [Non-textual]
Term Group: 1 Institution: Translation Centre Reliability: 1 [Adm.Inf.] Term: Banque de terminologie de l'UnionDeprecated: No Type: Term >Ref: Banque de terminologie de l'Union - Base de réflexion pour un projet de structure [Lookup Forms] Comment: Le projet IATE fusionnera toutes les données terminologiques qui ont été développées par les différentes institutions européennes en une base de données interinstitutionnelle unique afin de développer au maximum l'accessibilité des données, d'améliorer la qualité et de réduire les chevauchements d'efforts des différents groupes terminologiques [Validation] [Term Homs] [Word Level]	Term Group: 1 Institution: Translation Centre Reliability: 1 [Adm.Inf.] Term: Termbank of the Union Deprecated: No Type: Term [Lookup Forms] [Validation] [Term Homs] [Word Level]

Type: Document
Author(s): Jean-Marie Dufrasne
Title: Banque de terminologie de l'Union - Base de réflexion pour un projet de structure
Date: 2000
Publisher: -

Figure 3: From hit list to low level detail

INTERACTIVITY

The above mentioned ATOS feasibility study confirmed a common-place observation: the usability of a terminology database in the translation process cannot, or at least not exclusively, be expressed in number of entries stored in the database. Translators complaining that they cannot find a given new term in, for instance, Eurodicautom was reported as being a frequent phenomenon by the authors of the study. 1.4 million entries becomes a figure of purely academic virtue if you cannot find a valid solution for the *one* word that gives you a headache in the translation of an urgent document.

Two aspects play a role when it comes to unsatisfactory coverage in the existing databases. Either a specific subject domain is marginal in the activities of the Union and thus the need to generate systematic glossaries was never felt. Or new political and social questions come up, bring along with them a new vocabulary that needs yet to find its binding expression in all languages of the Union. The critical, uncertain, phase for the translators lies between the appearance of vocabulary in reality and the time when, once it is mastered, it has become common. As early as possible in this phase a terminology database should offer a solution to speed up this process.

The ATOS study clearly analysed a lack of inter-activity in the terminology arrangements of the institutions as the main obstacle that prevents the terminology production cycle from being faster. In many cases valuable terminology work done by translators in the course of their daily work remains unknown to their colleagues, as most databases do not allow direct write access for a larger population of users. Often terminology is hidden in private MultiTerm databases or waits on the "to do" lists of a few privileged colleagues who actually have the right to add something to a general database.

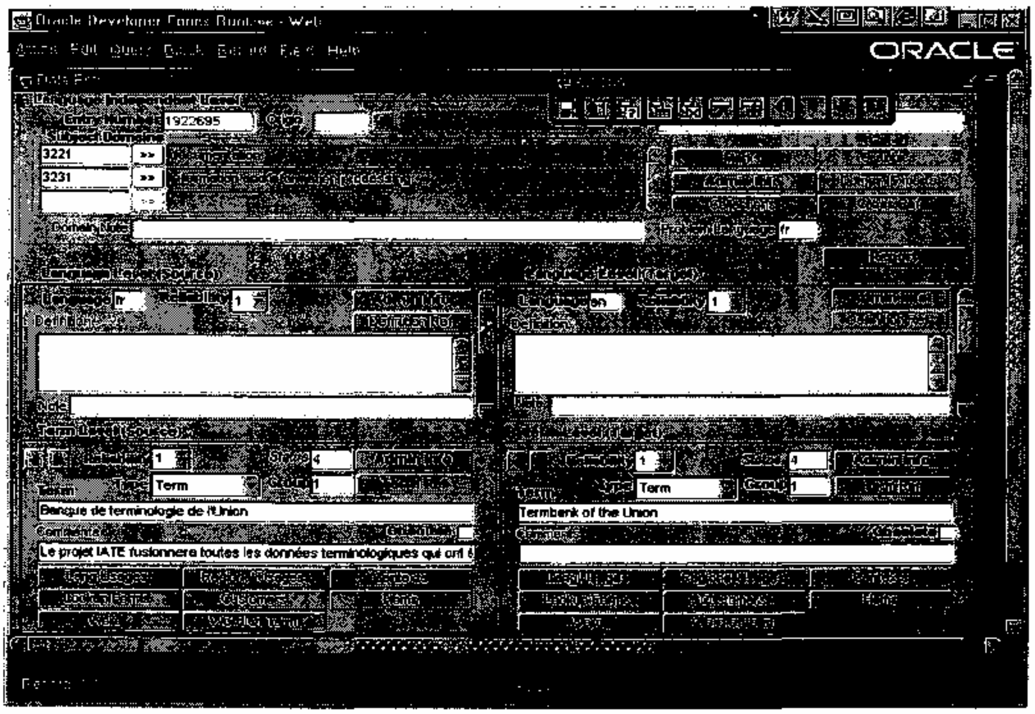
It was not only technical limitations of the early database systems that made the people in charge of the terminology resources of the Union reluctant to grant write access too freely to colleagues who are - although language experts - not necessarily trained terminologists. It was also the fear that if everybody can contribute directly and unfiltered to a terminological collection chaos will break loose. Given an easy user interface people may well abandon the paper glossaries, hand written cards etc. to make the results of their reflections available to their colleagues immediately. More pertinently, terminology would be circulating and give valuable aid in the day-to-day work of the Institution's translators. Still - what about the reliability of the translations proposed? What about the completeness of the terminological entries created this way? What about a certain ideal of terminological quality and coherence that should not be easily dismissed as "academic"? The key question for each system that chooses interactive feeding by a large population of contributors is how to ensure a certain quality standard in the data collected and published. In our case: how can we avoid creating a huge, uncontrollable interinstitutional terminology *scratchpad* that might satisfy some ad-hoc needs, rather than a reliable database that a wider, non-professional public can turn to in confidence?

However there need not be a clash of terminologists vs. translators - the underlining question is how to reconcile two potentially opposing requirements: the pragmatic need to disseminate information as quickly as possible to avoid redundancy of work, to make colleagues aware that someone has already taken care of a problem, and – more than that -- to what extend a problem has been solved. Terminology that is created by non-experts as problems arise may be fraught with a number of shortcomings: time constraints may simply not make if possible to provide complete documentation for a new term, the specific knowledge necessary to create a complete terminological entry may be lacking - just think of the labyrinthine complexity of some domain classification systems to see the point.

Question such as the above had already been discussed in the ATOS study. Two interinstitutional workgroups dealing with the integration of the term base into the workflow of the different institutions and the problem of data validation helped to develop a number of strategies that should make it possible to reconcile the need to produce terminology ever faster and the requirements of high quality standards.

RICHNESS VS. COMPLEXITY

The backbone of each terminology database is its data structure; it defines the degree of detail and complexity the database allows to maintain. And it may well be the first obstacle to efficient interactive data entry by non-expert users.



Entry Number 1922695

3221

3231

Problem Language fr

Language fr Reliability 1

Language en Reliability 1

1 Stage 4

Type Term Group 1

Term Banque de terminologie de l'Union

Comments Le projet IATE fusionnera toutes les données terminologiques qui ont é

1 Stage 4

Type Term Group 1

Term Termbank of the Union

Figure 4: Data entry screen

The brief and deliberately incomplete description of the data structure of IATE given in the above figure is an excellent illustration of the dilemma mentioned above: this structure definitely caters for the creation of very complete, self sufficient terminological entries – but who will ever have the time and the know-how to fill in all the information this structure could hold? The problem also has a very practical side to it: how can a user interface present all these possibilities in a user friendly way - i.e. without scaring the translator's away from the product and thus reducing the notion of inter-activity to a purely theoretical status?

A modern database system offers of course quite a number of features that can assist users in the phase of data entry: a rather small sub-set of the data structure will be defined as mandatory and will thus be presented in a user interface accordingly, i.e. mandatory information will be made easily accessible on the interface whereas more exotic elements will be hidden in sub-screens; the system will check on the presence of these mandatory fields to avoid incomplete information being accidentally stored. Where appropriate lists of closed value-sets will be used to avoid inconsistent usage of attributes. An interinstitutional work group is in the process of comparing the writing rules for terminological entries that have been developed in the different institutions. The result of this work will, where possible, lead to new automatic checks and to the addition of an on-line help system that will give valuable hints to non-expert users for each type of information that can be entered.

But what would be the information that is considered *mandatory* in this context? When a new terminological entry has been created it should fulfil two requirements: it should be meaningful for other users of the database who search for information on a given term. And it should contain sufficient elements to allow somebody to evaluate and if necessary improve the quality of the information given. The evident elements that spring to one's mind for the mandatory fields are: domain, language, the term itself, the source of the term and an example of its usage.

VALIDATION WORKFLOW

The above already indicates that from the outset of the project it had been envisaged to integrate procedures that would support the review of new or modified terminology. This meant basically supplying technical solutions for the formalisation of the proofreading of terminology. But it goes beyond the good practice of having a new entry checked by a colleague: the concept of a "validation workflow" was developed that would organise the cooperation of different actors (translators, linguists, terminologist and domain experts) in the terminology production cycle. The process would take into consideration the specific competencies of the people involved and would cater for a review of terminological entries on different levels: spelling, content, coherence, exhaustiveness etc.

In an early phase of the project a two level validation workflow was foreseen: The first one would be an internal review: the validation mechanism would route an new entry to another member of the same institution for an initial check; once the entry had passed this stage it would be sent, in the second phase, to a pool of domain experts from all participating institutions and, possibly, external organisations. This approach was rejected as some institutions wished to maintain complete control over their data and would not accept validation by others; it became also clear that a fixed two-stage approach would not be suitable for all institutions.

Today the EU term base offers a fairly flexible system of validation that allows for the definition of different validation cycles for each participating institution whilst not ruling out the option of interinstitutional cooperation in this field. A validation cycle is the sequence of validation stages. The number of stages, the actors of each stage and the type of checks that they should perform can be defined by each institution. An example will help to make the basic idea clearer: a simple validation cycle could contain the following three stages:

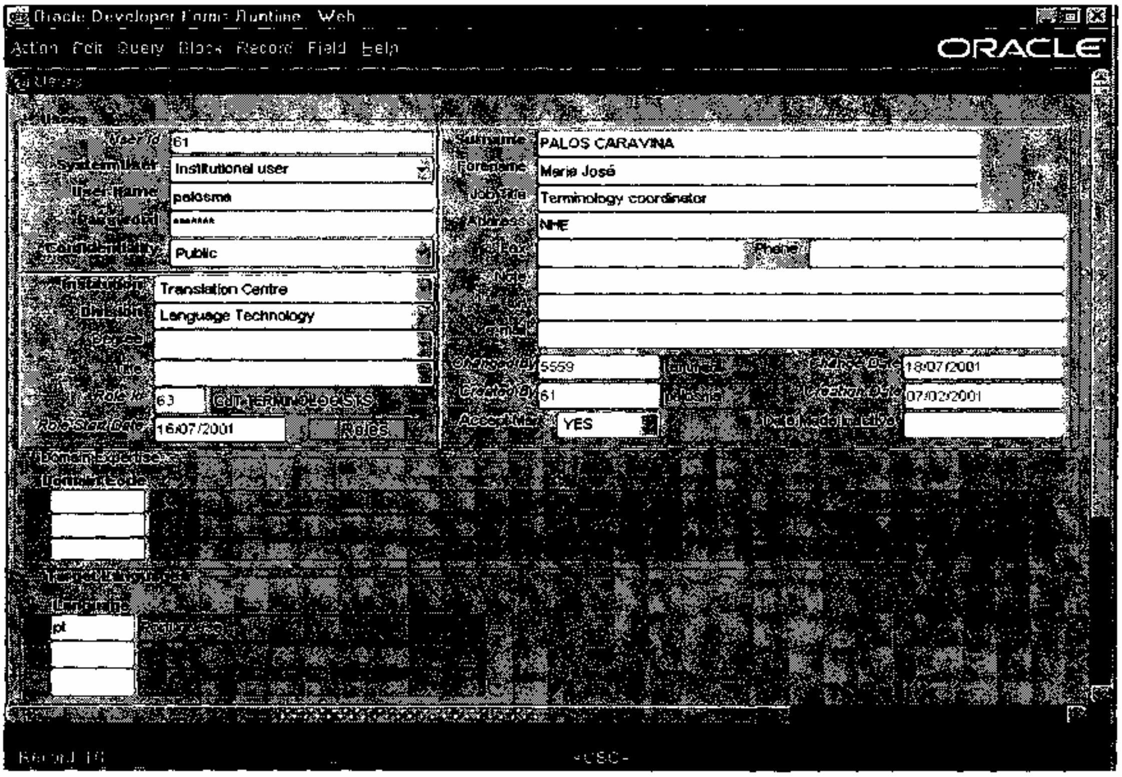
- Stage 1: formal check. This stage is launched directly after the creation or modification of an entry. The system will send the entry to a colleague who has the competencies to check its formal correctness, i.e. the spelling.
- Stage 2: content check. Once the formal check is accomplished the entry will be routed to a domain expert who will verify the contents and enrich it when appropriate.
- Stage 3: final check. A final coherence check by a terminologist terminates the validation process.

The system makes it possible to add validation phases (up to nine at the moment) if necessary but also to reduce the validation cycle to a single stage. The recommendation of the work group on this question was of course to have a least one validation stage. As each institution is free to handle this question according to their needs and possibilities it was necessary to introduce an element that would guarantee a certain coherence. A common set of validation statuses was thus defined that links the different institutional practises to each other and indicates to the users the degree and type of validation a terminological entry has undergone. As an entry will be visible, i.e. retrievable, even when it has not yet been validated this information is an essential indicator for the assessment of the entry's reliability.

This sequential approach to validation, that aims to benefit as much as possible from the competencies of a large population of participants, has the obvious advantage of providing the basis for a thorough, in-depth review of data. On the other hand it may hold the risk of creating an unbearable administrative overhead on terminology co-ordinators and thus turn what is supposed to be a work*flow* into a dead end. To avoid a situation where hundreds of entries remain non-validated a strategy had to be developed that would allow for the automatic distribution of validation work to the appropriate people. This strategy had to take into account the fact that many actions in the validation cycle are closely linked to language competencies and domain expertise, i.e. attributes that are specific to individual users of the database.

This kind of information will be maintained in the interinstitutional database for each user. This "user profile" also contains administrative information like the user's name, e-mail address, postal address, password, the institution the user works for etc. The fact that each user is known to the database system allows also for keeping track of individual preferences for the various activities that the database offers, as for example the search language or the sorting criteria for query results.

The essential element of the user profile for the validation process is the user's *role*. Roles offer the opportunity to group different users with common characteristics together. All users belonging to the same role share the same access rights to the system, i.e. they are allowed to perform the same type of actions. The rights for certain roles can be very restricted; the role "Guest" could, for example, only grant read access to the database. Other roles, like "Translator", "Expert Translator", "Terminologist", "Domain Expert", could make if possible to add, modify or delete entries. Roles also have an impact on access to the various sub-systems of the database. The right to launch specific administrative modules, e.g. large scale data export or import, is governed by the user roles. Here again the underlining concept in the implementation of this functionality was to provide flexibility for the needs of the participating institutions. Each partner of the project is free to define the roles they consider necessary for their organisation.


Oracle Developer Forms Runtime - Web
Action Edit Query Block Record Field Help
ORACLE
Roles
Phone

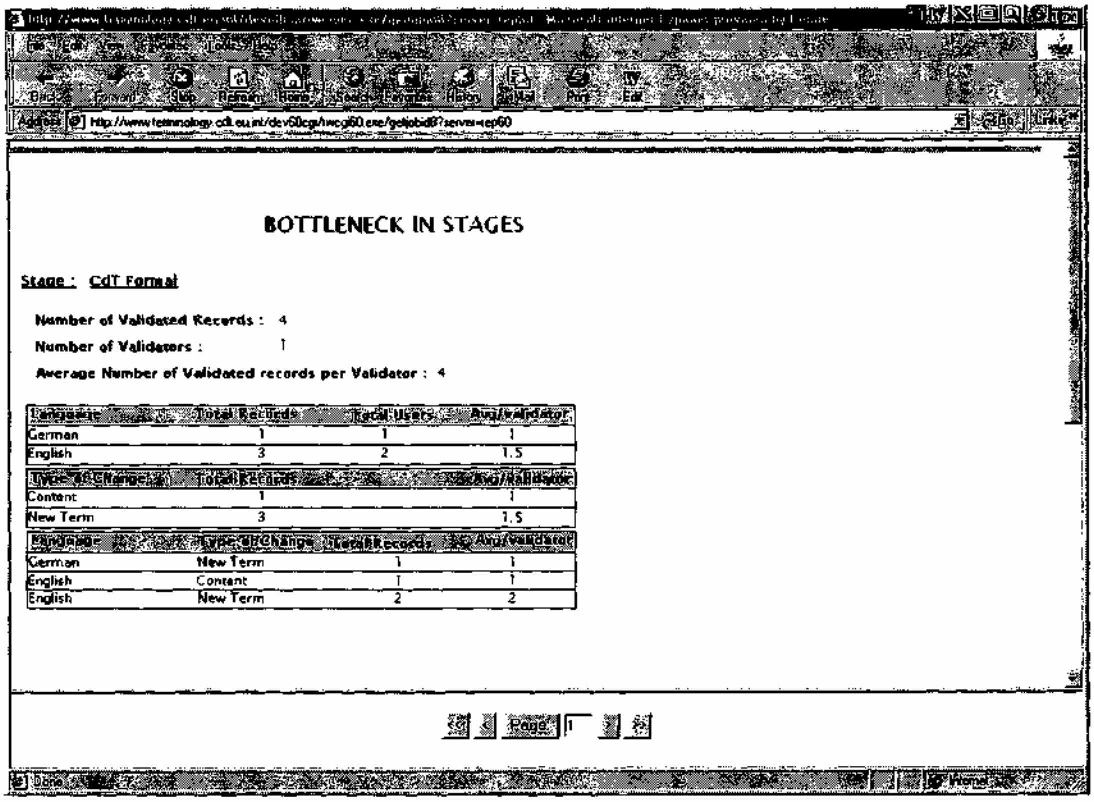
User Id 61
System User Institutional user
User Name palosma
Password ******
Confidentiality Public
Institution Translation Centre
Division Language Technology
Role Id 63 CdT TERMINOLOGISTS
Role Start Date 16/07/2001
Surname PALOS CARAVINA
Forename Marie José
Job Title Terminology coordinator
Address NHE
5559
18/07/2001
Created By 61
07/02/2001
Accept Mail YES
Domain Expertise
Domain Code
Target Languages
Language
pt

Figure 5: User management in the EU term base

The combination of individual user profile and the general role the user is assigned to is used to manage the validation process. The stages in the validation cycle are associated with specific roles, but they may also depend on language competence or domain expertise. Based on this information the system can distribute work to a suitable validator for each new or modified entry. The role of the author of a term is also taken into consideration for the triggering of validation cycles: it determines whether a complete validation cycle has to be performed, or if the role that users belong to allows to reduce the number of stages.

Based on the experience with the existing databases we can assume that a few thousand entries will be added or modified each month. Given the amount of validation work that this might imply each user of the database should also be an actor in the validation process. The system provides a simple on-line user interface, an "inbox", that displays a list of terminological entries that have been assigned to users of a specific role. The list contains information on the type of changes that have triggered the validation process, e.g. "new term", "formal change", "content change". The validates can use the information to prioritise their work.

The potential complexity of the validation workflow – just keep in mind that the system allows for a flexible set-up of all the elements involved - made it necessary to provide a number of tools that would help system administrators to monitor the process and to intervene if problems occur. Such problems could be the disruption of the workflow if no user with the required competencies can be found. The various reports make it possible to monitor the following parameters: validation work per validator and stage, comparison of two stages or cycles, bottlenecks in validation stages or cycles and dead ends in the validation process for specific entries. Dead ends and bottlenecks can thus be detected and managed. A specific interface allows administrators at any time to change the assignments of the system manually.



http://www.terminology.cdt.eu.int/dev60cgi/iwcgi60.exe/getjobid8?server=iep60

BOTTLENECK IN STAGES

Stage : CdT Formal

Number of Validated Records : 4

Number of Validators : 1

Average Number of Validated records per Validator : 4

Language	Total Records	Total Users	Avg/Validator
German	1	1	1
English	3	2	1.5

Type of Change	Total Records	Avg/Validator
Content	1	1
New Term	3	1.5

Language	Type of Change	Total Records	Avg/Validator
German	New Term	1	1
English	Content	1	1
English	New Term	2	2

Figure 6: Example of a report on a bottleneck in a validation stage

COMMUNICATION MECHANISMS

Validation as it is implemented in the EU term base is a strongly formalised way of cooperation between colleagues. A specific event - the creation or modification of an entry in the database - triggers a pre-defined sequence of stages that lead to a clearly defined goal: attribution of the label "Finally validated" to the entry in question. Another kind of cooperation, less formalised, one might even say deliberately open to improvisation, is the direct communication between users of the database. A database user might come across entries that he or she wishes to comment upon. This comments can be extremely useful if they are directed to the right persons. The IATE system uses so called "marks" to support this kind of activity. Marks can be attached to each entry and can be send to individual users or users groups (e.g. the terminology group of a specific language division) - again the system takes advantage of the information stored in user profiles and role definitions to simplify this task. Usually the contents of the marks will be an exchange of information and opinions. It could be the information that two specific entries, that represent the same concept, should be merged. Or that an entry is lacking essential information.

As long as a mark has not been removed by a competent colleague - i.e. once the described problem has been fixed - the mark text will be visible to all users of the database and inform them on ongoing work or help them evaluate the suitability of a entry for the problem they are working on. Besides the marks the database also offers an internal messaging system that can be used to communicate problems of a more general nature - i.e. comments that are not related to single entries - to other users of the database.

REPORTS AND AUDITING

Besides the above mentioned reports on the status of the validation process the EU term base will offer a considerable number of other monitoring tools to help administrators with the task of managing the database. These reports include statistics on the work activity of users belonging to specific roles or institutions. Tools that make it possible to extract statistical information on the growth and the current state of the database have also been integrated. Finally, basic operational statistics on the usage of the system can easily be created.

A complete audit trail on the linguistic information stored in the EU database makes it possible to follow-up on the modifications carried out on each entry and - if necessary - to restore previous versions. The auditing records the type and content of a modification and keeps also information on the user(s) of the database who have changed an entry.

CONCLUSION

Perhaps the best words to sum up the different strategies used in the EU term base to ensure both quality of the terminological data and more efficient integration of terminology work into the translation workflow are communication and cooperation: the former by "showing" a modified entry to other users (as in validation) or by providing the technical facilities to make sure that comments end up with the right people; the latter by offering a platform that allows actors of the same or different bodies to share their competencies. Although there is still some way to go until the database will actually be accessible to the general public, both within and outside the institutions, the prototype has already shown that the database of the Union is indeed becoming a reality.

Although, the implementation of the system has, from a technical point of view, already reached a rather advanced stage it is predictable that user feedback during the prototype and pilot phases will still lead to a number of changes. The biggest challenge of the project, however, lies in its introduction in the terminology and translation workflow of the participating bodies. This is illustrated in the second part of the paper by the example of the European Parliament's Translation Service.

PART TWO - A CASE STUDY: ADAPTING TERMINOLOGY ORGANISATION AND PRACTICE AT THE EUROPEAN PARLIAMENT TO IATE

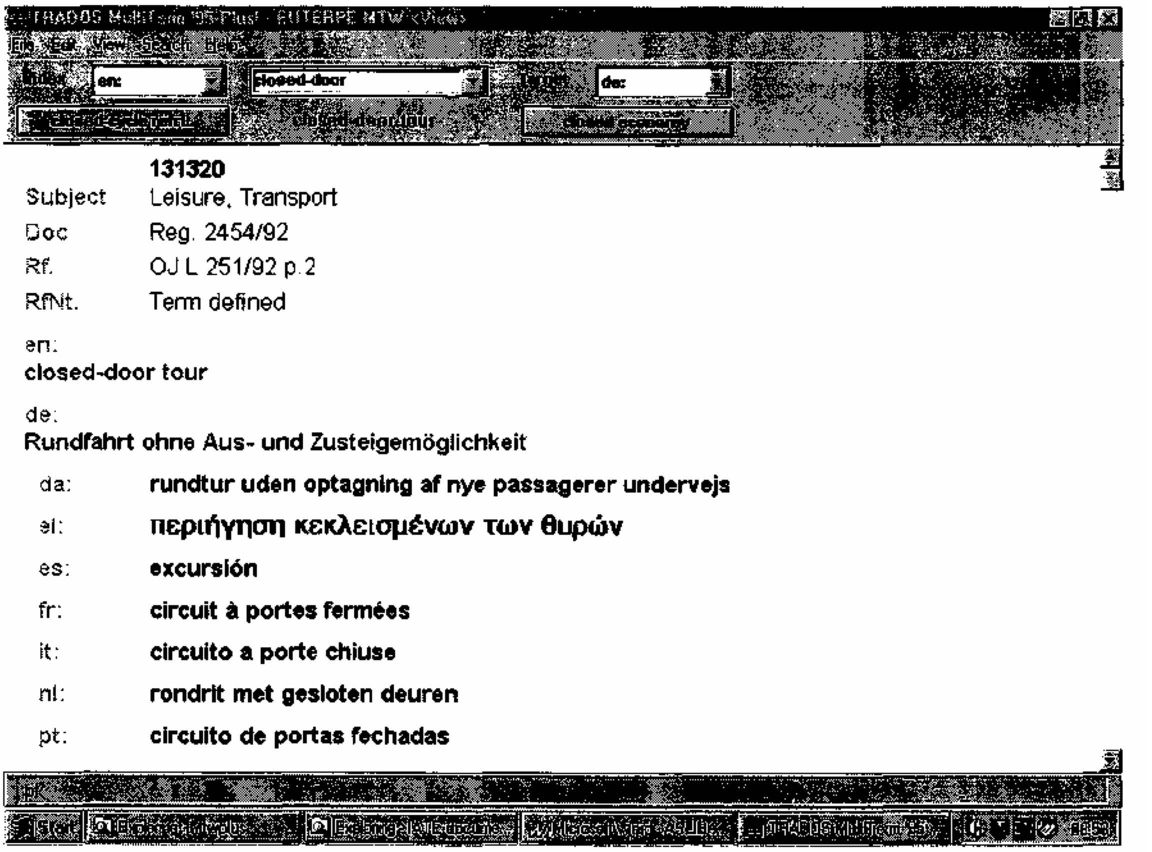
This part of the paper will describe the current terminology scene at the EP, the attractions of the IATE project for us and the challenges of adapting both to the IATE structure as envisaged and to changes in terminology practice. Section 1 will describe current practice, section 2, the strategic attractions of IATE for the Translation Service, section 3, the attraction of IATE for EP translators and terminologists, section 4 will give a brief overview of our problems with IATE to date and section 5 will present some conclusions and thoughts for the future.

1.CURRENT TERMINOLOGY ORGANISATION AND PRACTICE AT THE EP

The EP has had a terminology service in some form since the 1960s. After many years of organisation on traditional lines as an independent unit with terminologists for all official languages researching and publishing thematic glossaries and an in-house journal, it was reorganised in the early 1990s as one part of a larger department called the SILD Division (for *Division du Support informatique, linguistique et documentaire* or "IT, Language and Documentation Support Division"). The title refers intentionally to "language support" as opposed to terminology alone since for a number of years now activities have also included other support services for translation, i.e. the introduction of and on-going support for the Trados Translator's Workbench, text alignment, translation memories, speech recognition and a Parliament-wide document-production system (DocEP).

The current terminology team comprises five terminologists (almost all of whom also have other tasks as part of their job description) and one secretary. It is still responsible not only for managing but also for initiating almost all terminology activity within the translation directorate, although a number of translation divisions act as service providers for languages not covered by the SILD team and translation divisions occasionally initiate terminology projects. The main on-going project covering all 11 EU languages involves monitoring and logging the terminology used in the Official Journal of the European Communities (OJ), which is an ideal source for EU translators since it exists in all languages in parallel text and is most often the terminology that we are obliged to use. To take account of other needs, since the EP is an essentially political institution, we also collect and collate topical terminology in an attempt to anticipate our translators' queries.

All terminology work, whether in SILD or translation divisions, is carried out in MultiTerm 95 database management software, which is compatible with the Translator's Workbench, with a data structure adapted to our particular needs. To provide greater flexibility we are about to offer translation and other staff not familiar with MultiTerm the possibility of supplying terms via a simple intranet form for further processing by a terminologist. Terminology records, whether created in MultiTerm or the new interface, are included in the EP terminology database, EUTERPE (for *Exploitation unifiée de la terminologie au Parlement européen* or "European Parliament one-stop terminology management system") which currently contains approximately 260000 records in some or all of the EU languages, with acronyms or abbreviations where relevant, and sometimes with Latin (for scientific terms) or non-EU languages (for political parties, national or regional institutions, etc.). A typical EUTERPE entry from the OJ gives subject-domain information, a reference to the document where it occurred and the publication details for the OJ, information about whether the concept is defined in the OJ and terms in all languages.



	131320
Subject	Leisure, Transport
Doc	Reg. 2454/92
Rf.	OJ L 251/92 p.2
RfNt.	Term defined

en:
closed-door tour

de:
Rundfahrt ohne Aus- und Zusteigemöglichkeit

da:	**rundtur uden optagning af nye passagerer undervejs**
el:	**περιήγηση κεκλεισμένων των θυρών**
es:	**excursión**
fr:	**circuit à portes fermées**
it:	**circuito a porte chiuse**
nl:	**rondrit met gesloten deuren**
pt:	**circuito de portas fechadas**

Figure 7: A typical EUTERPE entry

Providing the translator sets his own language as the target, this approach is designed to give all the relevant information at a glance, without cluttering the screen.

2. THE STRATEGIC ATTRACTION OF IATE FOR THE TRANSLATION SERVICE

As the first part of the paper has described, IATE was originally a Translation Centre project. However, it is safe to say that, from the point of view of the major EU institutions in general and the EP's Translation Service in particular, it was a project whose time had come. In the current climate where resources are scarce and a major enlargement of the EU is just over the horizon, it makes economic sense. How can anybody justify funding at least three major terminology databases and a number of smaller ones, when a single database could cover all the institutions' needs and, taking all the institutions together, cost less? I have to say "taking all the institutions together" because there are few direct financial savings for the EP, since MultiTerm is bundled with the Translator's Workbench available on all translation staff's PCs and almost all management of our database is the responsibility of in-house staff. However, if you look back at my example of the "closed-door tour" and take account of the fact that on

Eurodicautom there is not one but a number of entries for that self-same concept all from the same source, the scope for reducing duplication of effort is obvious.

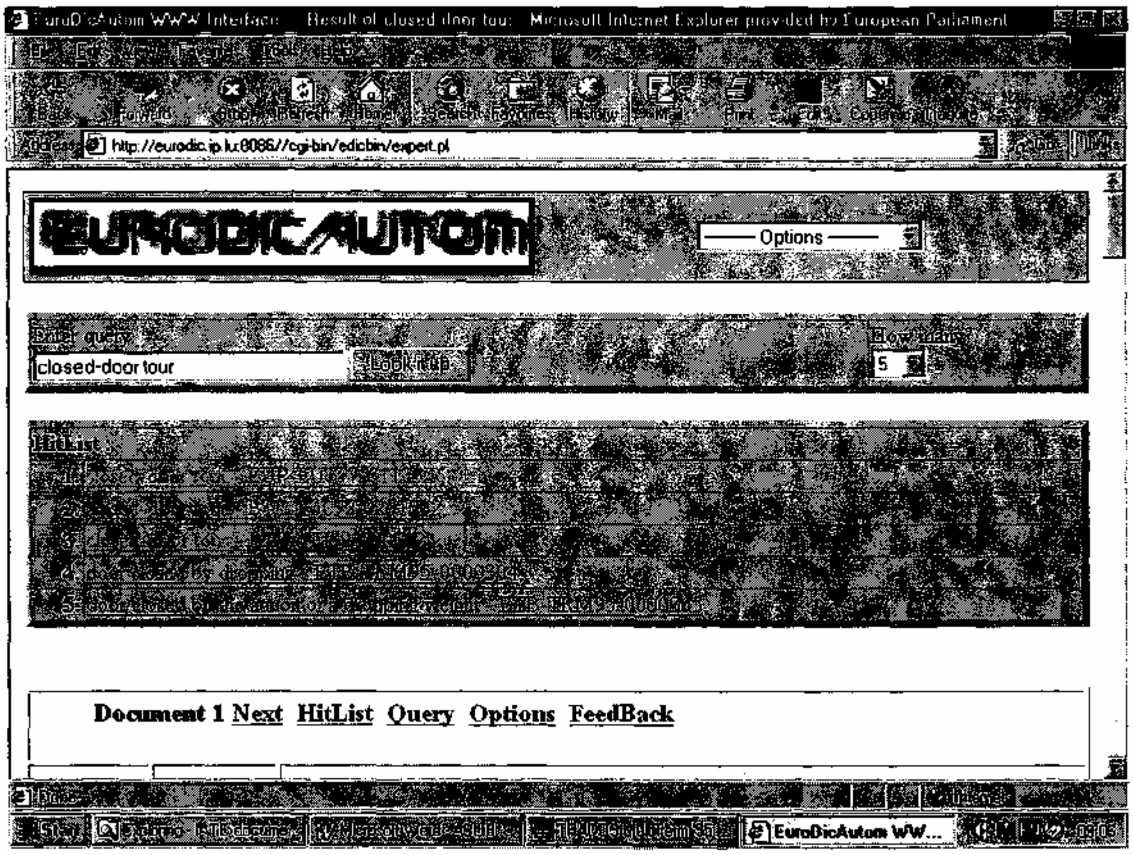

EuroDicAutom WWW Interface - Result of closed door tour - Microsoft Internet Explorer provided by European Parliament
http://eurodic.ip.lu:8086//cgi-bin/edicbin/expert.pl
EURODICAUTOM
Options
Enter query
closed-door tour
Look it up
How many
5
HitList :
Document 1 Next HitList Query Options FeedBack
EuroDicAutom WW...

Figure 8: Entries for closed-door tour on Eurodicautom

In a world where each institution works quite independently, results like the one above are understandable but they remain regrettable, nonetheless, particularly since TIS, too, includes a record for the same concept.

Moreover, as I will show shortly, there should be savings in staffing terms for the EP as well as for the other institutions and, in an era when every post counts because we are gearing up from 11 languages to perhaps 21, that is a very significant economic incentive. This is not to suggest that our senior management is opposed to terminology activity as such or views terminologists as unproductive because it is more difficult to quantify their productivity than it is for translators. Indeed, raising the profile of terminologists and public awareness of their activity and effectiveness is seen as one of the major positive points of the whole exercise.

3. THE ATTRACTION OF IATE FOR EP TRANSLATORS AND TERMINOLOGISTS

For our translators - at least those who combine terminology activity with translation, or would like to - the attraction of the IATE database is that it is designed from the outset as an interactive system. For a number of years we have had sufficient problems with MultiTerm in our specific environment to restrict access to EUTERPE to the core team within SILD. All other translators consult a fixed copy of the database (updated regularly) and either write to buffer databases from which terminology is taken over into EUTERPE by SILD staff or e-mail their proposals to us (or as I said before, they can use the new intranet terminology form). They find this unsatisfactory because they, often rightly, view us as too slow to react and, with a small team with many other responsibilities, we find it difficult to keep up with the workload and verify some of the changes proposed.

Once IATE goes live all our translators will, in principle, be able to propose new terminology records in their working languages with equivalents, if relevant in their mother tongue. They will do their work directly in the single database and all institutional users of the database will have immediate access to it. The first check on the correctness of a new record will be made by a reviser or senior translator with the same mother tongue and an expert knowledge of the language from which the source term came. A terminologist will then intervene to mark the new record for the attention of other divisions who ought to add their languages and, perhaps, where the subject domain lies outside the EP's realm of competence, an outside expert who can provide concept-level validation. As the final stage of validation the terminologist will check that the record complies with IATE standards in general and, except in the rare event of information being confidential, it will then be on general release. The flow chart illustrates how this should work in practice.

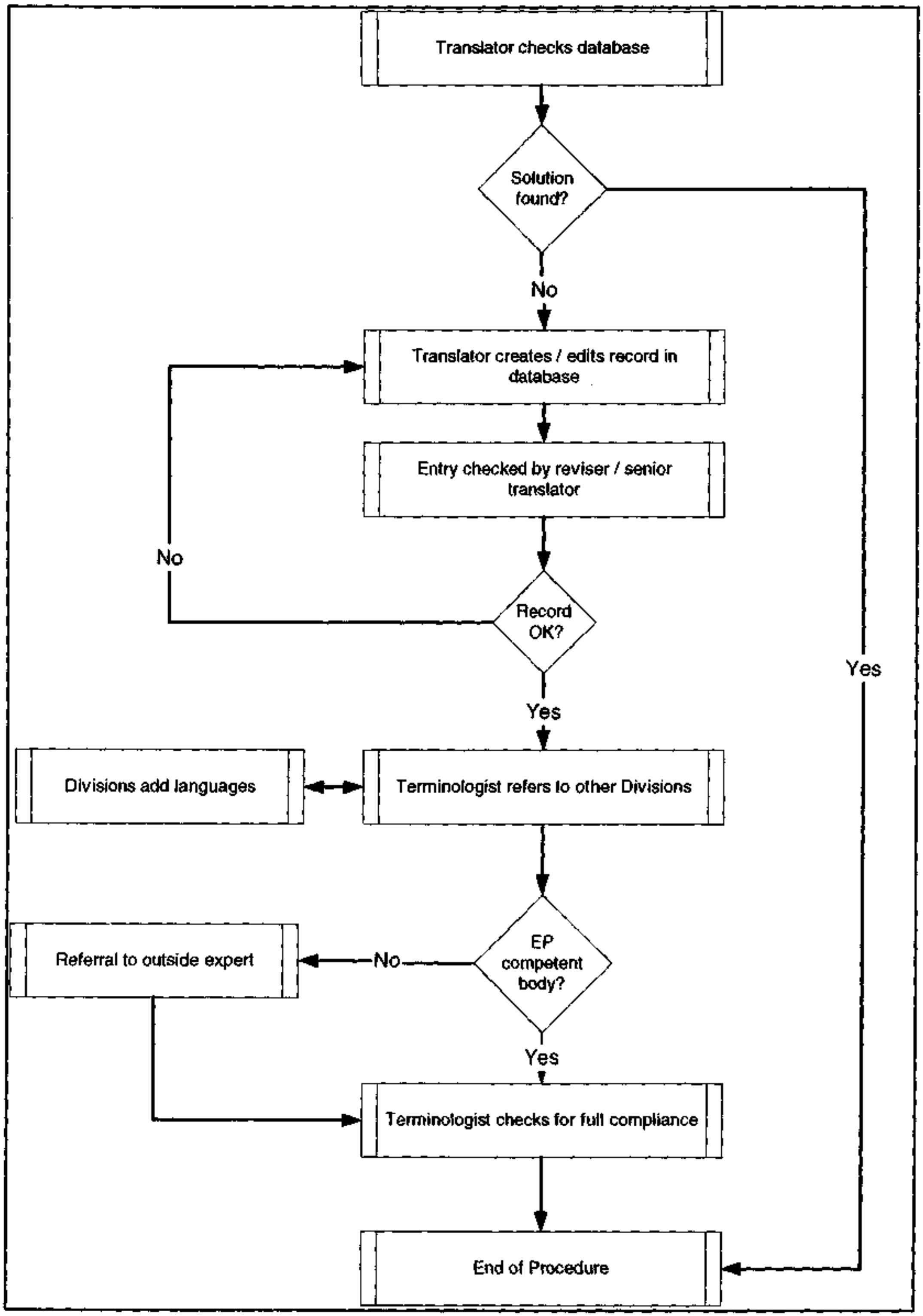

Translator checks database
Solution found?
No
Yes
Translator creates / edits record in database
Entry checked by reviser / senior translator
No
Record OK?
Yes
Divisions add languages
Terminologist refers to other Divisions
Referral to outside expert
No
EP competent body?
Yes
Terminologist checks for full compliance
End of Procedure

Figure 9: The IATE input and validation cycle as envisaged at the EP

From both the translator's and the terminologist's point of view this will push terminology activity to an earlier stage in the translation cycle, making it more useful to the translation community while ensuring that, over time, any ad hoc solutions which are less than satisfactory are revised. (The IATE structure allows for terms to be marked as deprecated and thus not available to the general public.) At the EP, of course, since much of our work involves Commission proposals on which the Parliament and Council comment or take decisions after they have been discussed at Commission level and a certain amount of terminology established, we hope that we will also have timely access to work done in the other institutions as well, to eliminate duplication of the type illustrated earlier. As for the Official Journal terminology project, this should be taken over by the IATE central management, using state-of-the-art terminology extraction software.

Furthermore, an approach of this type is essentially based in translation divisions with a central unit required only for coordination and harmonisation. It is therefore very economical in terms of staff and resources, allowing for scaling up in both languages and areas of activity without an increase in the number of terminologists in the central unit.

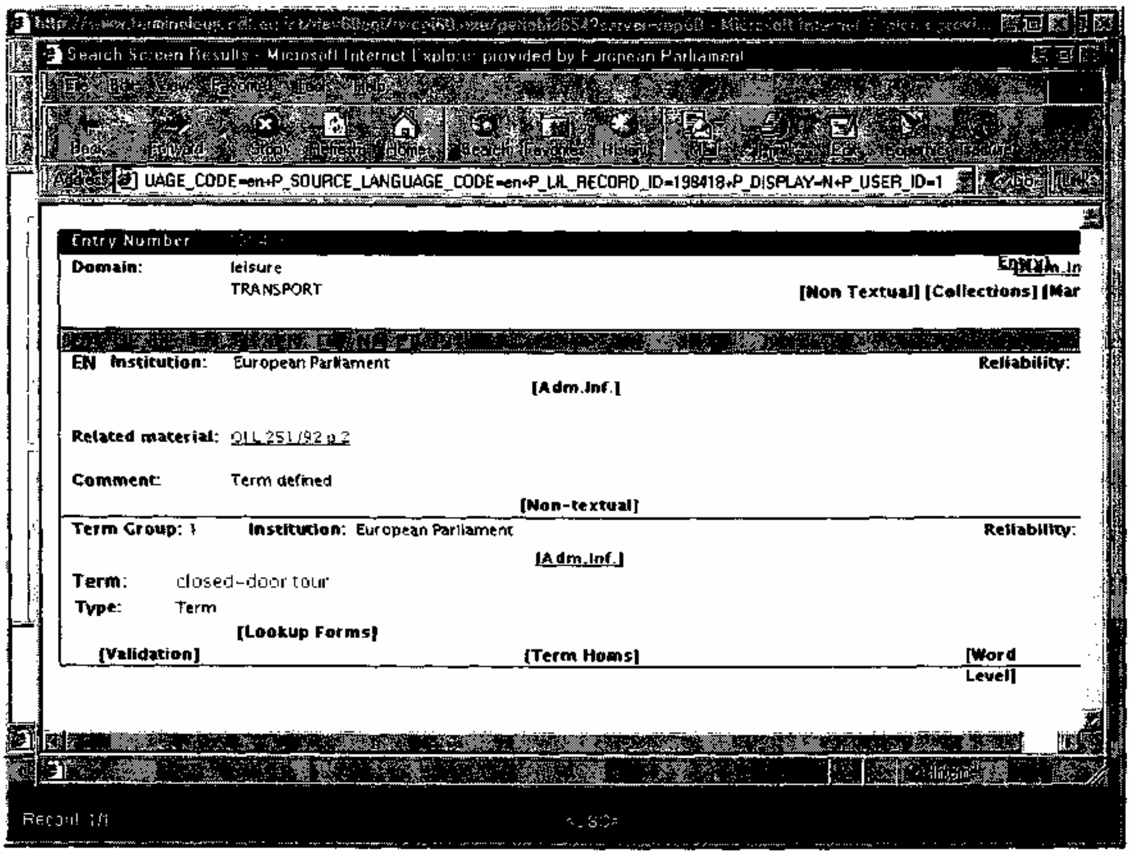
Since the IATE system includes two forms of communication, one for general use and one for use between terminologists, we should also be able to improve communication with translation divisions involved in terminology activity and feedback from other terminology users. The e-mail system will be available to all users, although we hope that they will restrict it to terminology issues. Of course, since the system is parametrically defined, if it is abused we can request the administrator to suspend users' rights.

The other communication system, known as "marks" and described in part one, concerns records rather than users alone. All users will be able to read marks attached to records, so that if they see something incomplete they will be made aware that updating is on-going, but terminology coordinators and terminologists will also be able to create them and, eventually, to delete them, once action has been taken. When they log on to the system they will be informed of the number and type of marks marked for the attention of their unit, so that they can prioritise their work and, if necessary, distribute it among colleagues to reduce response time.

It will also be possible to address marks to external collaborators. You may recall that, in commenting on the input, revision and validation cycle foreseen, I referred earlier to the opportunity of external validation for terms outside the EP's area of competence. This is something for which we have never had the resources in the past, although we have all taken (often unfair) advantage of our friends, families and acquaintances with specialist knowledge on occasion. One of the aims of IATE is to build up a network of national and international experts able to validate information and provide input for their field, which will then be accessible to all database users. This will allow us to structure best practice, to everybody's benefit.

4. IATE PROBLEMS AND CHALLENGES

So far, so good, but as with all interinstitutional projects, not everything is wonderful, at least not yet. Since we are only at the prototype phase, it would be wrong to insist on aesthetic shortcomings, although a brief look at part of my standard example in IATE format gives an idea of what I mean.



Search Screen Results - Microsoft Internet Explorer provided by European Parliament

UAGE_CODE=en+P_SOURCE_LANGUAGE_CODE=en+P_LIL_RECORD_ID=198418+P_DISPLAY=N+P_USER_ID=1

Entry Number

Domain: leisure
TRANSPORT

[Non Textual] [Collections] [Mar

EN Institution: European Parliament Reliability:

[Adm.Inf.]

Related material: OJ L 251/92 p 2

Comment: Term defined

[Non-textual]

Term Group: 1 Institution: European Parliament Reliability:

[Adm.Inf.]

Term: closed-door tour

Type: Term

[Lookup Forms]

[Validation] [Term Homs] [Word Level]

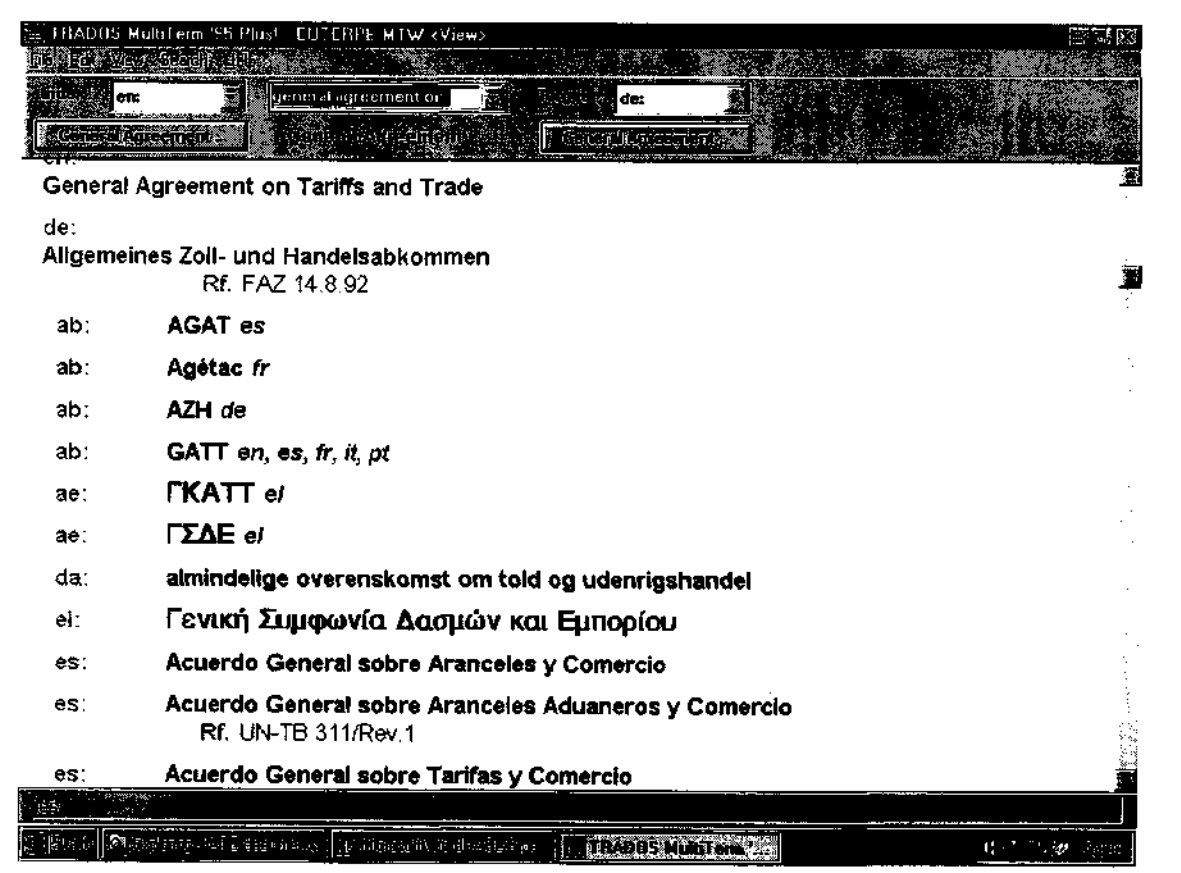
Record 1/1

Figure 10: Part of the IATE bilingual display of "closed-door tour"

Far more difficult for us, at least at the present time, is coming to grips with the changes in approach required. One of the advantages of having a very small team to manage EUTERPE in MultiTerm has been administrative simplicity. For hands-on terminologists it is rather daunting to change over to a situation where, quite simply, you are so dependent on other system users, even for something so easily centrally managed as deletion of redundant entries or the merge function, which works very simply in MultiTerm but, at the moment, seems much more complicated in IATE. However, being realistic, that would have to change too, quite soon, even in MultiTerm. With 21 languages no one terminologist would be able to recognise whether all terms were singular or plural, let alone whether they represented the same concept, so that merging and deletion would have to be reorganised.

We also have internal problems with the differences in structure between EUTERPE and IATE, some of which are dependent on MultiTerm 95 as such, some relate to our data structure and some to more profound differences of interpretation. For the problems of the first type which do not lend themselves to automated solutions (primarily the presence of multiple synonymous abbreviations and terms in any language for the same concept) we will have to request some type of "validation holiday" in the interim period between final loading of data and the system being regarded as live, to avoid triggering a validation avalanche or tsunami when we sort them into the correct term groups. Some of the problems of the second type will require the same treatment, primarily for terms and abbreviations from non-EU languages, but others will hopefully lend themselves to automated solutions once we have implemented changes to data presentation which are currently under assessment.

The last type is the most intractable, but is something which we ought to have addressed long ago. Quite simply, in our data structure where abbreviations are entered in separate indexes (for Latin and Greek characters) and not with terms, we have allowed users to obscure the difference between an abbreviation being used in a particular language (or even at the lowest level occurring in a text in a particular language) and being a term which belongs to the language concerned.



en:
General Agreement on Tariffs and Trade

de:
Allgemeines Zoll- und Handelsabkommen
Rf. FAZ 14.8.92

ab: AGAT *es*
ab: Agétac *fr*
ab: AZH *de*
ab: GATT *en, es, fr, it, pt*
ae: ΓΚΑΤΤ *el*
ae: ΓΣΔΕ *el*
da: almindelige overenskomst om told og udenrigshandel
ei: Γενική Συμφωνία Δασμών και Εμπορίου
es: Acuerdo General sobre Aranceles y Comercio
es: Acuerdo General sobre Aranceles Aduaneros y Comercio
Rf. UN-TB 311/Rev.1
es: Acuerdo General sobre Tarifas y Comercio

Figure 11: A EUTERPE entry with multiple abbreviations

In this example we would have no less than five Spanish terms once the conversion to the IATE structure is complete. Would that be the right solution? The practice came about partly because of a major shortcoming of MultiTerm, which does not lend itself readily to cross-language searching and partly because, in an environment where Translation has little control over input, we wanted to ensure that as many possible searches as possible for obscure abbreviations would be successful. They usually are, but I am sure that the terminology actors at the EP who work on data consolidation will have years of work as a result.

However, all of these problems pale almost into insignificance when compared with that of getting the IATE system up and running in time. The time plan (kick-off meeting to live system in 171/2 months) was always hopelessly optimistic, but the project is now way behind schedule and we have to hope that there is no further significant loss of time if we are to have a new system bedded in by the next enlargement of the EU. The new terminology input and validation model requires cooperation from all existing divisions to work efficiently. They will need experience before new languages are added.

5. CONCLUSIONS

We remain optimistic that we will solve our problems in time to allow our current terminology users to gain the necessary experience of the IATE system of cooperative working towards common goals before enlargement. We are confident that once they have learned to use the full functionality of the new system they will see it as a vast improvement over the past. Moreover, we are sure that the availability of all our terminology throughout the EU and beyond will be appreciated by translators everywhere.

Does this mean that, once the development phase of the project has successfully been accomplished by mid 2002, the future of the terminology in the EU will be all bright, i.e. free of shortcomings like inadequate coverage or disappointing quality? Unfortunately the EU term base alone will not do this magic trick. The system that is being developed at the moment comprises a number of features that go beyond what is "state of the art" in the field today. But independently of the features that the system will offer, it remains only a tool. It may be powerful, it will hopefully be user friendly, but it will definitely be most efficient if used by well-trained colleagues who see systematic terminology work, both the creation and validation of new concepts, as part of their profession. This approach, that was also a recommendation of the ATOS study, demands reinforcement of training efforts and a wider awareness of the crucial place terminology holds in the working cultures of the institutions.

But then again, the EU term base could well become more than a tool. It will hopefully become a vehicle that will promote the idea of interinstitutional cooperation in the field of terminology. The discussions in various working parties of the IATE project in the last few month show that the enthusiasm for such cooperation is clearly increasing.

References

Almeida, A. (2001), 'Die terminologische Datenbank der Europäischen Union' in *Festschrift Joachim Göschel, Beiträge zu Linguistik und Phonetik*, Ed. Angelika Braun, Supplement to the *Zeitschrift für Dialektologie und Linguistik,* Stuttgart

Ball, S.I. (1993), 'The European Parliament's Euterpe Database: An Introduction' in *TKE 93: Terminology and Knowledge Engineering,* Cologne, Index Verlag, pp. 308-315

Ball, S.I. (1996), 'In the Beginning Was the Glossary: the development of integrated language support services at the European Parliament' (Paper presented at *Au commencement était le terme : la terminologie au service des entreprises, EII*, Mons, Belgium)

Ball, S.I. (2001), Terminology Activity at the European Parliament' in *EAFT Update* March-April 2001

European Commission, Translation Centre (1999), *IATE – Services for the Development of an Interactive Terminology Database System, Open Call for Tenders* DGIII/99/050-IDA-101.02/01/IATE1, Luxembourg/Brussels

Johnson, I., Palos-Caravina, M.-J. (2000), 'Validation and Quality Control Issues in a new Web-Based, Interactive Terminology Database for the Institutions and Agencies of the European Union' in *Translating and the computer 22,* Aslib, London

Johnson, I. and MacPhail, A.(2000), *IATE - Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union,* Workshop on Terminology Resources and Computation, LREC 2000 Conference, Athens, Greece

MacPhail, A. (2000), *IATE - Inter-Agency Terminology Exchange,* Conference for a Terminology Infrastructure in Europe, Paris, France

Quality and Reliability S.A. (2000), *IATE Project Validation Work Group Proposal,* Athens, Greece

Vidick J-L. and Defrise C. (1999), *Interinstitutional Terminology Database: Feasibility Study,* Atos, Brussels, Belgium, 147 pp.

List of Abbreviations

CST: Centre for Speech Technology
CdT: Translation Centre for bodies of the European Union
CIT: Interinstitutional Committee for Translation
EP: European Parliament
EU: European Union
EuroVoc: European Vocabulary
EUTERPE: European Parliament one-stop terminology management system
LATE: Inter-Agency Terminology Exchange
OJ: Official Journal of the European Communities
SILD: Information-Technology, Language and Documentation Support Division
TIS: Terminological Information System (Council of the EU)