[*Translating and the Computer 22*: Proceedings of the Twenty-second international conference…16-17 November 2000. (London: Aslib, 2000)]

Validation and Quality Control Issues in a new Web-Based, Interactive Terminology Database for the Institutions and Agencies of the European Union

Ian Johnson and Maria-José Palos Caravina

Centre de Traduction des Organes de L'Union Européenne
1, rue du Fort Thüngen
L-1499 Luxembourg Kirchberg
{Ian.Johnson, Maria-José.Palos_Caravina}@cdt.eu.int

Abstract

We present an ongoing project for the creation of a single central terminology database for all the institutions, agencies and other bodies of the European Union. The background, objectives, benefits and main features of the system are briefly introduced, followed by a presentation of the solutions proposed to resolve the complex validation issues addressed by a project which involves interaction between many institutions with different internal validation processes as well as access from the general public.

Background to the IATE Project

The aim of the IATE project is to create a single central terminology database for all the institutions, agencies and other bodies of the European Union which will provide translators and terminologists with a centralised source of EU terminology data. The project, which is funded by IDA (the EU programme supporting the Interchange of Data between Administrations), started in January 2000 and the first phase - the detailed design and specification of the system - was completed in July 2000. The implementation of the prototype system is currently in progress and is scheduled to be installed for user testing by the end of November 2000.

Three expert groups comprising representatives of all institutions as well as some external experts from the member states worked on the technical specifications relating to the data structure, validation process, and workflow integration requirements of the new inter-institutional terminology database, the results of which were summarised in the project's first major deliverable, the System Analysis and Design document. The work of these groups having been completed, an implementation support group was set up to provide guidance and feedback to the contractors throughout the development phase of the project. This group will be particularly concerned with user interface design, evaluation of the results of the data uploading process and further issues relating to the content of the database.

Data will be imported from existing institutional databases, principally Eurodicautom (European Commission), Euterpe (European Parliament), and TIS (Council of the European Union), as well as from some smaller databases developed at a number of other institutions and agencies including the Court of Justice, the Court of Auditors, the European Social Committee/Committee of the Regions, the European Investment Bank and the Translation Centre of the Organs of the European Union. This data uploading process will necessitate the development of tools to detect duplicate entries and to assist in the process of merging overlapping entries. Once the final import of data into the new database has been made, the other databases will be discontinued and the efforts of the institutions will be combined to maintain and develop the single, centralised IATE database. The validation process has been designed to be flexible enough to handle the widely differing processes existing at present, as well as the proposed inter-institutional validation aspect anticipated in the future. The database will be integrated into the principal components of the translation workflow of the various institutions, particularly word processing (MS Word). In September 2000, extended funding was granted by IDA to support further consolidation of the data and integration of the system with additional tools used by the institutions, including especially Trados Translator's Work Bench and Word Perfect, as well as the Glossary Production System (currently used at the Council), the Thesaurus (currently used at the Court of Justice), the One-Stop Shop (currently under development at the Commission's Translation Service) and the Trademark Translation Workflow System (integrating Translation Memory, MT and workflow, currently being developed by the Translation Centre). Further funding may be requested at a later stage to integrate LATE with even more systems, such as the Commission's Systran MT system and Euramis, the translation memory and document retrieval system.

The next steps in the project are to implement the prototype (by November/December 2000) and test it in two pilot phases with increasing numbers of users at the various institutions. The system is scheduled to be completed by July 2001.

Validation and Quality Control Issues

The objectives of the validation part of the terminology database design and development have been to set up formal acceptance rules that during interactive input control automatically whether an entry shall be accepted or refused, to design content-related access schemes in connection with the definition of access rights for validation staff, and to work out administrative procedures to ensure that participating institutions or bodies can cooperate in the validation process. The original proposal in the Call for Tender was for a two-stage validation workflow. The first stage would be an internal review whereby new data would be first routed to other members of the same organisation for checking before being distributed for central validation according to domain and language combination to a pool of domain experts selected from the staff of all the participating organisation and possibly also other organisations. However, it became clear in the course of the analysis phase that a number of participating organisations wished to maintain independent control of their own data and were not currently in a position to handle validation by external experts. It therefore proved necessary to design a flexible system capable of handling both types of validation which would permit these institutions initially to continue to use their current internal validation process whilst at the same time providing the means allowing them to move gradually towards a standardised, inter-institutional process in due course.

Validation Workflow Cycles

In order to cater for the differing validation workflows that exist in the different institutions participating in the project, it has been necessary to design a flexible and dynamic workflow model which can easily be adapted to the particular (and changing) processes of each organisation, whilst at the same time providing the structures necessary for gradual inter-institutional cooperation. Institutions must define the point at which they wish to release their data to public view and define the number, type and sequence of internal validation stages they require. Also, it is necessary to define the different validation cycle for different types of users (e.g. translators, terminologists, language/domain experts, system administrators, etc.). It is felt that this approach offers a clearer and easier gradual integration of the validation cycles of the participating institutions and agencies than an alternative approach which was considered (according to which each institution maintains one validation cycle consisting of a fixed sequence of stages and which users join at the stage specified for their role). Such a process requires the specification of the validation status of each stage in the cycle, i.e. the visibility of the term, how 'fixed' the term is, the user role required to perform this stage, and whether specific language/domain knowledge and/or institution membership is necessary for the stage. Users in each institution are grouped into different roles which are defined and maintained by the institution's administrator. Each role will be associated with different access rights (e.g. read, insert, update, delete, merge, export, import, change validation status, user and role maintenance tool, insert/delete marks, etc.). Information on individual users (e.g. name, password, source language(s), target language(s), domain expertise, role, institution, division, contact coordinates, start/end date of role, etc.) will also be maintained, together with a profile of the user's preferences (e.g. search language, search result sorting criteria, list of display languages for search results, etc.).

Data Entry Validation

The process of validation of an entry starts from data entry and continues through to final validation. The system has been designed to support users during data entry with easily accessible displays of the rules that are applicable to a given entry. Where possible automatic checks to verify data entry are carried out which have been derived by comparing sets of rules used in the different institutions. A complete audit trail showing all changes to any entry in the database is available off-line to the system administrator.

Other features of the entry such as context information to provide either an example of the occurrence of that term or the authority/reliability of that term or confidentiality (rarely of the term as a whole, more usually of specific fields such as source and references) are provided for in the database and checked in the validation process. The ability to enter/modify reliability codes will only be assigned to validators. The range of reliability values that will be available to the system at language level and at term level will be as follows:

Code	Meaning	Explanation
0	information at term or language level has been downgraded prior to deletion	this code is system-imposed if the information has been merged automatically onto another record and hence is now duplicated in the database; when no automatic merge has been performed, this code can only be given by an authorised user from the division that owns the information, pending deletion of the entire record
1	information at term or language level was entered by a non-native speaker	this code is system imposed and can only be upgraded by a native speaker
2	information at term or language level was entered by a native speaker	this code is system-imposed and can only be upgraded by an authorised user
3	information at term or language is well-documented, the code having been upgraded	this code has been given by an authorised user, when he/she is confident that the information is well documented

During data entry, the system will provide a default value of 2 if the language of the entry matches the target language(s) of the author, otherwise 1. When a term or concept has been marked for deletion, the reliability value will automatically drop to 0. As far as confidentiality values are concerned there will be the following three possible confidentiality scopes: Public, All Owners of Database, Institution. The first indicates that the data is not confidential, the second that it is confidential to the public, and the last that it is confidential to all outside a particular institution. Confidentiality will apply either to whole entries, a specific language of an entry, human and/or documentary reference sources and comments. Some institutions are already developing translations for terms in languages of those states which may join the EU in the near future, but for political reasons these terms can only be used internally at present and must therefore remain confidential to the public. Other agencies undertake work on behalf of security services or in a legal context and it is necessary that some of the information relating to an entry be kept confidential.

The system will automatically detect duplicate entries in cases where there is a 100% match. We will also evaluate strategies for dealing with entries which are very similar but not exact matches (in order to check whether the entries are, for example, spelling or inflectional variants). In cases where duplicate records exist with translations in languages which do not overlap, it is difficult to define a straightforward automatic detection procedure without the use of a pivot language. However, since the vast majority of source terms is in English or French, the number of non-overlapping duplicate entries is not expected to be very large. In fact, as new translations *are* added to existing terms, many non-overlapping duplicate entries will eventually overlap, at which point the system will propose that they be merged.

Validation of New Proposals, Interactive Updates and Ownership of Data

We distinguish between classes of users who have 'direct write access' to adding new data and those who have 'deferred write access' to propose additions to the database. In the case of the latter, the proposals will be visible only to the author of the proposal and to the people who evaluate them. Such proposals will not enter the validation process until they have been examined and accepted by the data managers. The data managers will be automatically notified when new proposals have been sent to them and they will be provided with tools which enable them to prioritise the proposals according to certain criteria depending on the type of change concerned (e.g. spelling correction, translation of a given term in new languages, or completely new record, etc.) Once a data manager has accepted a proposal, the new entry will enter the validation process and become visible to all users in the institution. The aim of this process is to reduce the overhead on validators. A separate database structure for the maintenance of tables of new proposals is provided. Translators can specify the person they wish to evaluate their proposal via a validation assignment tool.

As far as interactive updates are concerned, modifications to an existing entry in a particular language are not allowed at any level while that entry is in the process of being validated. Instead, users may use so-called working fields or marks to make any proposed changes known to the validators. In this way it is up to the validators to review any such proposals and to decide what to do with them. It may be that the proposed change requires the validator to return to an earlier stage in the validation process - this is achieved by returning the validation status code to a lower value. In order to prevent the validation process becoming overloaded, it is only modifications to certain key data fields that trigger the re-validation process, in contrast to such modifications to existing data for a particular language entry, the addition of new data (e.g. term synonyms or translations in further languages) is not restricted by the validation process and proceeds in the usual way. Users and validators (with the appropriate access rights) may only modify data in their target language(s). The management of interactive insertions and deletions is similarly controlled through the mechanism of roles and access rights, with rules restricting the ability to add or delete data while parts of the term are being validated.

When users make changes to existing data or propose new entries, the system prompts them to indicate the type of change made. This helps validators to sort and prioritise the work they receive in their inbox. Each inbox entry will also include information concerning the date of the validation request, the language of the term(s), the list of term subjects, the term(s) and the name of the author or modifier. In the case where a term is owned by a different institution to the one in which the validator works, it will be necessary for the validator to contact the institution concerned and explain the changes they need to make. It is advisable to restrict the ability to change the subject code of a term - which affects all translations of the term - to the final validation stage which could be language independent.

The question of the ownership of the data has an impact on the interactive update and input process as well as the validation process. The institution that first created or introduced the concept becomes the institution responsible for the entire entry. Users from other institutions are permitted to insert new translations and synonyms to existing terms. However, if they wish to propose a change to an existing term, this must be done by making the suggestion(s) in the working fields of the entry and notify the last validator, modifier or creator of that entry. The ability to delete entries is similarly governed by roles and access rights. However, a user who has been given the delete right may not delete any data his/her institution does not own, nor any data other than his/her target language, nor any data that is being validated. As this rather restricts the effectiveness of this right, it was decided that a user will be able to at least initiate the deletion process for any of the data structure levels owned by his/her institution. If the data of other institutions is going to be affected by this action, the system will automatically provide the user with the coordinates of those who need to be notified. Only after all institutions involved have signalled their agreement to the deletion proposal, will the deletion be carried out.

Validation Process Monitoring Mechanism

A monitoring mechanism will also be provided in order to draw to the attention of the system administrator any problems which might arise in the validation process and enable the settings to be adjusted to improve the performance of the system. Such problems include disruption to the validation flow because no user profile matches the validation criteria for a particular term or because there is some mistake in the validation flow settings, a dead end to the validation flow because the validator is absent for a long time or has left the institution, or a bottleneck to the validation flow caused by a particular validator being overloaded with validation work. The validation process monitoring tool will help the administrator or data manager select a new validation cycle or stage and validator in order to bypass such problems.

Illustration of the Validation Process in Action

In order to provide an illustration of the way in which the validation will work, let us suppose we have the following users, each with different language skills and domain knowledge, and each based in a particular division of their institution:

User Name	Source Language	Target Language	Domains	Role	Institution	Division
User 1	English French	Greek	IT Finance	Terminologist	Institution Z	Division 1
User 2	English German	Greek	IT Law	Translator	Institution X	Division2
User 3	English German French	Greek French	IT	Experienced Translator	Institution Y	Division 3
User 4	English German	Greek	IT Finance	Internal Validator	Institution X	Division 4

This user data is entered by the local system administrator. Each user is assigned a role in the validation process defined for that particular institution. These roles may of course vary from institution to institution, but a maximum number of 9 is currently permitted. The roles are associated with different validation cycles and database access rights:

Role Description	**Validation Cycle**	**Access Rights**	**Institution**
Translator	Cycle 1	Read Write Update	Institution X
Terminologist	Cycle 3	Read Write Update Delete Merge	Institution Z
Internal Validator	Cycle 2	Read Write Update	Institution X
Expert Translator	Cycle 2	Read Write Update	Institution Y

The different validation cycles consist of an ordered sequence of validation stages, as exemplified below:

Cycle Description	**Validation Stages**
Cycle 1	1,2,3
Cycle 2	2,3
Cycle 3	3

The validation process is divided into a number of stages at which a series of checks are carried out by users associated with specific roles and a validation status is assigned to the entry. The validation status code determines whether the term is visible to other users and whether it is fixed. The validation status values are used consistently across the different institutions, but this does not necessarily mean that all have to use three validation stages - institutions that need more will be able to define sub-stages for validation stages 1 and 2. An initial code O is used to indicate that no validation has taken place. Code A means that the form of an entry has been checked for compliance with data entry and writing rules, including spelling. Code B means that the content has been validated, and finally code C means that validations A and B have both been performed:

Validation Stage ID	**Stage Description**	**Validation Status**	**Role**	**Language Specific**	**Domain Specific**	**Institution Specific**
1	Formal Check	A (not visible, not fixed)	Internal Validator	Y	N	Y
2	Domain Language Check	B (visible, fixed)	Experienced Translator	Y	Y	N
3	Final Check	C (visible, fixed)	Terminologist	Y	Y	N

Taking as our first example an English IT term entered by User 2 who has Greek as source language and is based at Division 2 in Institution X, the validation process runs as follows. As User 2 is a translator, the process starts at the first stage of validation cycle 1. This stage requires a validator who in this case must have source language English and target language Greek, and be located in the same institution. The system directs the validation to User 4 who satisfies these criteria. In the case of multiple results the system will allocate the term to the most available validator according to a set of criteria defined by the system administrator.

After validation by User 4, the entry is assigned validation status A and passed on to the next stage of the process - in this case, stage 2 which requires an experienced translator with source language English, target language Greek and who is a specialist in the IT domain, but who need not be at the same institution as User 2 or User 4. The system identifies User 3 as having the corresponding requirements and passes the entry along for validation. When complete the validation status code of the entry is updated to B and the entry is passed on to the third and final stage in the validation process. This stage requires a terminologist with source language English, target language Greek and domain expertise in IT. In our example, User 1 has the requisite attributes and therefore receives the entry for final validation. Once validated, the validation status of the entry is updated to C and the validation cycle is now complete.

Of course, communication between validators (and indeed others, if necessary) is possible at any stage in the process via the system of marks which are stored in the database attached to the terms they relate to and which enable database users (excluding the public) to pass on comments and questions to other participants in the process. Insertion and deletion of marks will be controlled by the access rights mechanism, which will allow all database users the right to insert marks but only validators will be permitted to delete them. Certain information about the author of the mark will be recorded automatically by the system (e.g. institution, role, date of creation). Further information may be provided by the user concerning the language to which the mark refers and the type of mark (e.g. validation, deletion, merging, or general). On accessing a term in the database, users will immediately see whether there are any marks associated with the entry and will be able to filter them according to the author's name, their institution and role, the language the mark refers to, the type of mark, the date of creation, and the name of the person to whom the mark is addressed. A general report tool allowing sorting on the same criteria will be provided to enable users to locate marks irrespective of database entries. In order to allow a user to send a message to a particular person directly (e.g. if the proposed new entry or modification needs to be examined urgently, or if there is a need for confidentiality), a messaging system will be provided, but in most cases it is expected that the system of marks will be the preferred mechanism for communicating between users validators and administrators of the database.

Summary

We have presented issues concerning the process of validation and quality control in a web-based, interactive inter-institutional database. The need to incorporate the different validation cycles and practices currently used at the different institutions has required us to design highly flexible and parameterisable validation workflow processes and tools. The movement among EU terminology groups towards inter-institutional cooperation addressed in this project points to the growing need for the standardisation of terminology data representation and the harmonisation of terminology production and validation processes.

Acknowledgements

The authors gratefully acknowledge the contribution of the LATE project sub-contractors Q&R (particularly Fotis Zografos, Dionysis Ninios and Yota Asimakopoulou) as well as that of all members of the IATE project technical working groups in producing the Systems Analysis and Design document version 4.0, upon which this paper is based.

References

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, Luxemburg/Brussels

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 *System Analysis and Design v4.0,* Athens, Greece

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