IATE – Inter-Agency Terminology Exchange: Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union

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

Alastair MACPHAIL
Centre de Traduction des Organes de L'Union Européenne
1, rue du Fort Thüngen
L-1499 Luxembourg Kirchberg
Alastair.MacPhail@cec.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 planned system are introduced, followed by a review of the issues being addressed by the three technical groups working in the areas of data structure specification, validation procedures and workflow integration.

Introduction

The aim of this communication is to 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. We intend to start with a very brief presentation of the Translation Centre for the Bodies of the European Union, which is the service responsible for implementing this project, before going on to explain the history of the project, its administrative context and schedule. We will continue with a general overview of the current state of terminology activities in the EU organisations and the issues that the project aims to address, before going on to explain the objectives and expected benefits of the project and the main features of the planned system. The project is currently in the analysis and design phase, and the full specification of the system is in the process of being defined. The second half of the paper will review some of the main issues under discussion in the technical working groups responsible for defining the functionalities of the system, with a view to illustrating the overall scope and aims of the project.

1 Centre de Traduction (CdT)

The Translation Centre for the Bodies of the European Union is a self-financing decentralised agency of the European Union set up in 1995. Its primary function is to provide translation services on a fee-paying basis to the EU organisations. Another of the CdT's missions is to promote cost-savings in areas in which there is duplication of effort between the translation services of the institutions and it is in this context that the IATE project was launched.

2 History

The EU institutions have been discussing the possibility of merging their terminology databases for many years. The first practical step in this direction was taken in 1998 when the Interinstitutional Translation Committee (ITC), the body responsible for interinstitutional cooperation in the area of translation, terminology and documentation, gave the Translation Centre a remit to commission a feasibility study into the creation of an interinstitutional database.

The contract was awarded to the Brussels IT consultancy firm ATOS and the study was conducted by Mr Jean-Luc Vidick and Ms Christine Defrise of the Université Libre de Bruxelles. Work started in September 1998 and the final report was delivered in March 1999. The main recommendations were:

1. An interinstitutional database is both technical feasible and functionally desirable
2. All existing data should be merged into a single database
3. A common data model should be adopted
4. Common rules for data presentation and evaluation should be defined
5. Cooperative management mechanisms should be established
6. Full interactivity for data input and updating

The report was adopted by the ITC in May 1999, but no decision was reached at that meeting on putting its recommendations into practice. The CdT, under pressure from its customers to provide adequate terminology facilities, decided, in the absence of any immediate prospect of an interinstitutional database being set up, to go ahead with its plans for an interactive web-based terminology facility to serve the requirements of the decentralised agencies. Having secured a promise of financing from DG Enterprise under the IDA programme, an EU programme for promoting the electronic exchange of data between the EU institutions and agencies and member states administrations, the CdT started drafting specifications for the new system. In July 1999, DG Enterprise launched a call for tenders on behalf of the Translation Centre under the title of the IATE project, which stands for Inter-Agency Terminology Exchange. When news of the CdT's initiative broke, the other EU institutions showed immediate interest, and at the ITC meeting of September 1999, it was decided that all the EU institutions would participate in the IATE project, transforming it from an interagency to an interinstitutional project.

The tenders were evaluated in September and in November the contract was awarded to the Greek IT firm Quality and Reliability (Q&R) with the Danish government research institute Center for Sprogteknologi (CST) providing the linguistic expertise.

The project schedule is as follows:

Stage	**Time frame**
Analysis and design	01/00-06/00
Development	06/00-12/00
Pilot I	12/00-03/01
Pilot II	03/01-06/01
Full production system	07/01

3 Current situation

At present, the big three institutions (the European Parliament, Council and Commission) have their own on-line terminology databases (Euterpe, TIS, and Eurodicautom, respectively), now available on the web. Some smaller institutions (European Investment Bank, Court of Auditors, CdT) have internal databases, generally using MultiTerm. Others make do with glossaries in WP formats, card files, etc. Certain institutions (European Social Committee/Committee of the Regions) have no systematic terminology arrangements.

Partial solutions have been developed to the problem of providing access to the full range of EU terminology resources. The contents of Euterpe and TIS are periodically uploaded into Eurodicautom, though the difficulty of the operation means that it is not done on a very regular basis. Moreover it is a relatively brutal operation and there is some loss of information (problem of field mappings and domain correspondence) as data from the other databases must be squeezed into the Eurodicautom format, and there is no consolidation of information from the three databases. The Commission has developed "one-stop" access to Eurodicautom plus range of terminology resources available on the web, including Euterpe and TIS. This is a very useful tool, but is only available on the Commission's internal website. There is some attempt to exchange terminology data between the other institutions, but generally speaking their terminology resources remain internal. The following three subsections summarise the key features of the major EU databases.

3.1 Eurodicautom (Commission)

- Coverage: 11 EU official languages plus Latin
- over 1.240.000 entries (c. 5 million terms) and 325 000 abbreviations and acronyms (July 1999)
- Domain classification : Lenoch Universal Classification (LUC)
- on-line and batch consultation (via Euramis)
- Client interface and web interfaces on intranet/internet (http://www.eurodic.echo.lu)
- Fed from work of Terminology Unit (based in Brussels and Luxembourg), contributions from translators systematised by Eurodicautom team and contributions supplied under contract by private companies and field experts
- Update weekly
- Translators have access to unit-level MultiTerm databases (110)
- Low interactivity (Eurodicautom updated weekly, data from unit-level databases filters through only very slowly into main database)
- Currently on BS2000, migration to UNIX/Oracle ongoing

3.2 TIS (Council)

- Coverage: 11 EU languages plus Latin and Irish.
- 200 000 records (600 000 terms) 45% contain 3+ languages. 25 000 records contain 5+ languages
- terminology reflects translation problems that have arisen in Council texts. 170 subject codes

- Feeding: New terms entered directly by Council's terminologists (5/6 per language division). Other users may enter comments or suggestions. New data searchable immediately (text search on all fields)
- Growth rate: 4 000 translations per month.
- IT system: Client-server application using FULCRUM SearchServer running under AIX on a Bull Escala server.
  PC-based user interface.
  Web interface: http://www.tis.consilium.eu.int

3.3 Euterpe (Parliament)

- Coverage : 11 EU official languages plus Latin
- 171 000 records
- bilingual entries: 92 887
  9-language entries: 13 556
  11-language entries: 5 623
- Some with notes or definitions or the Latin equivalent (botany and zoology) many with the corresponding abbreviations or acronyms
- A few terms (titles of political parties, ...) in non-EU languages
- The approach is descriptive rather than normative.
- IT platform: MultiTerm '95+ database, client interface, web interface on intra- and internet (http://www.muwi.trados.com/)

4 Drawbacks of current set-up

At present there is no single point of access to up-to-date terminology data of all the EU institutions. Interactivity is limited, except in the Council, so there is little user feedback, and the terminology cycle is relatively slow as users cannot quickly and easily suggest additions and changes to the data. There are problems of inconsistency in the use of terminology between the institutions and no easy way of standardising usage. There is no easy way of organising cooperation between terminology services of different institutions, with the result that there is considerable duplication of effort.

5 Project objectives & expected benefits

In order to address these shortcomings, the objectives of the project are:

- To provide a single point of access to all existing EU terminology resources
- To provide an infrastructure for the constitution, shared management and dissemination of terminology resources
- To provide a vehicle for the application of advanced language processing technology to terminology management
- To provide a basis for integrating terminology into the translation and document workflow
- To create a European platform for cooperation between EU institutions and terminology organisations in Member States

The expected benefits are:

- Faster access to wider range of terminology via single point of access to all EU terminology resources
- Enhanced terminology production by providing all actors in the terminology cycle (authors, translators, terminologists, domain experts) on-line access to a single database
- Shorter time-lapse between appearance of new term in "real world" and inclusion in database
- Elimination of duplication and redundancy of information
- More rational employment of human and financial resources
- Enhanced terminology quality through interactive validation procedures
- Greater user friendliness

6 Features of new system

Essentially, the new system will be a web application with a central relational database combined with a text search engine for speed of consultation. All interfaces will be web-based and accessible through any standard browser. Interactive data input entails fairly sophisticated management of user access rights and the number of participants and the different types of access to the system (consultation, data input, terminology validation and management) means that roles and profiles will have to be defined for the different user groups and user types. A key feature of the new system, and one that distinguishes it from most existing systems, is the integration of a validation workflow. That is to say that whenever new data is contributed to the system in the form of a new entry or an addition or change to an existing entry, it will be automatically routed to the person or persons responsible for checking and validating the entry. In this way it is hoped that opening the system up to all-comers will not lead to complete chaos. The number of participants in the system also means that management both of users and of data content will have to be distributed, so these functions will have to be accessible remotely.

7 Project organisation

As the project must satisfy the requirements of the IDA programme, which is financing the venture, and the requirements of all the participants a fairly complex organisation is required.

Overall monitoring of the project is the responsibility of an expert group called EGEUT (Expert Group for setting up an EU Terminology database). The group comprises a representative of each of the EU institutions, some of the decentralised agencies and member states. It reports both to the Interinstitutional Translation Committee and the Telecommunications between Administrations Committee (TAC) which is the governing body of the IDA programme.

Hands-on management of the project has been devolved to the Project Steering Group, which comprises a representative of each of the big institutions (EP, Council and Commission), plus the CdT and the IDA team.

A series of technical working groups have been set up to gather information and reach agreement between the participants on data structure, validation and integration with the document production and translation workflow.

8 Issues to be resolved

As we have said, the project is currently in the analysis and design phase, and the details of the system are still under discussion and have yet to be finalised. However, we would like to present the main issues that are being addressed by the three technical work groups in order to demonstrate the scope of the project and the complexity of some of the issues we are attempting to resolve.

8.1 Data structure

One result of earlier discussions on setting up an interinstitutional database was the elaboration of an interinstitutional data model, which has served as a basis for the current discussion. It is generally accepted that the new database should retain the conceptual model common to most existing data. The data structure which is currently being proposed is based on an analysis of the various databases held at the partner institutions and agencies, the IATE project feasibility study and call for tender, and the ‘Fiche Terminologique Interinstitutionelle’. Before some aspects of the structure can be finalised, however, it will be necessary to have the results of the Validation and Workflow Work Groups, so that the requirements for these can be taken into account. We also wish to ensure that the data structure defined for the database is as far as possible compatible with existing and emerging standards such as GENETER, MARTIF and SALT. To this end we have invited experts in these areas to join the Work Group.

In establishing a data structure for the inter-institutional database the most critical questions have been:

- The choice of subject code classification scheme (Lenoch, Eurovoc or other). Our current proposal is to use a slightly extended version of the Lenoch codes, since this is the system adopted by the largest amount of existing data. The subject domain picklist will be organised as a three-level hierarchy with short cuts and alphabetic display available as options. In order to provide the user with an overview of the codes, only the top level subjects will be shown when the hierarchy is opened; lower level subjects can be opened by the user.
- The internal structure of an entry. It is proposed to organise this in three levels: the language-independent level, the language level and the term level. Each entry covers a single concept.
- Whether it is possible to impose a pivot language or languages or a mere indication of the original source language is sufficient. Here we have decided that the question of imposing a pivot language is too sensitive a decision for the present and that in any case we currently have insufficient resources to implement this solution. So the specification of the original source language will have to suffice for the time being.
- Provision of the possibility of including lexical information where possible (i.e. information about part of speech, morpho-syntax and valency). This would be useful for development of spin-off language processing applications such as cross-lingual information retrieval.
- Hyperlinks to documentary databases(e.g. the Celex database of EU legislation) and fields for graphics and multimedia. Because such functionality is likely to be of increasing importance, it has been decided to include these in the design of the database right from the start, although they may not all be used immediately.

8.2 Workflow

An important concern in the integration of the system into the current workflow of the various institutions is to ensure that the translators and terminologists can continue to use it without needing to change the current workflow. This is because the workflow arrangements adopted by a particular institution or agency depend on many factors such as organisational structure, type of work, number of translators, availability of domain experts, deadlines, type/size/number/kind of documents, etc. The creation of a unique inter-institutional workflow is therefore not realistic.

The system needs to provide ways for communicating with system experts, domain/language experts and record creators/validators. This will be achieved by the use of email or the messaging system in the new database. Phone numbers will also be for immediate contact. This will enable translators who are working against tight deadlines to obtain critical information as quickly as possible.

Users of the database will need to be able to reference legal document sources (especially permanent public document servers such as CELEX). General writing rules will therefore have to be established for this purpose and hyperlinks provided to connect the user with CELEX and other such servers. Users will also be able to provide phrases, sentences and paragraphs as usage examples if they so desire. We have rejected the possibility of attaching the whole document to a particular term because it would cause storage problems and retrieval delays. Documentary references, usage examples and hyperlinks will be stored at the language level, with multiple values permitted as far as reference information and hyperlinks are concerned.

For purposes of maintenance and import/export of glossaries a batch retrieval function will be provided. This will allow the user to specify which records and fields should be included and also the format of the output file.

Integration of the database with a number of tools currently in use would be very useful. In particular, integration with word processing applications (such as Microsoft Word and Word Perfect) and computer-assisted translation systems (such as Translator's Workbench and Euramis) are regarded as high priority for almost all partners. The Commission's LDT Editor (for extracting portions of a database) may be used to provide the batch retrieval function to save reinventing the wheel. Integration with other tools such as voice recognition software, the Translation Centre's Trademark Workflow system and spell-checking tools, although highly desirable, will for the moment have to wait until the anticipated phase 2 of the project. We are also in the possibility of using pre-processing of a text by SYSTRAN to provide a list of candidate terms for the database. Such pre-processing could also present lists of terms found in the database to the translator as initial input to the process of translating the document. Certain agencies have also raised the possibility of integration between the terminology system and web search applications enabling multilingual search capabilities.

8.3 Validation

The objectives of the validation work group are 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 each participating institution or body be represented in the validation process.

The original proposal 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 appears that certain participating organisation wish to maintain complete control of their own data and can not accept validation by outsiders. It seems prudent to provide facilities for both forms of validation and see at the time of implementation which participants opt for which.

The process of validation of an entry starts from data entry and continues through to final validation. The system is being 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 will be carried out, and we are currently analysing current rules used in different institutions to determine whether a reliable set of rules which do not require complex processing can be established. 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. A complete audit trail showing all changes to any entry in the database will be available off-line to system administrator.

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.

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 adminstrators, 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, etc.). Information on individual users (e.g. name, password, source language(s), target language(s), domain expertise, role, institution, division, etc.) will also be maintained.

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.

Summary

We have presented the background and current work relating to the development of an inter-institutional terminology database for the EU. Important aspects of the design include integration with existing tools and workflow arrangements, definition of a data structure that can handle the variety of data maintained and is compatible with emerging standards, and the design of a flexible validation process that will allow institutions to work as they do currently but offer the flexibility for adopting new processes in the future. The design and specification stage of the database is now nearing completion and we expect to have finalised the details by June 2000 ready for the implementation of the first prototype by November/December. Although it would be wrong to underestimate the challenges to such an undertaking, the potential benefits to translators and the European public are enormous in terms of increased access to terminology data, ease of maintenance and extendability of the database and, last but by no means least, increased cost-effectiveness through the elimination of duplicated effort.

Acknowledgements

This work has been supported by funding under the IDA programme. Thanks are also due to many present and past members of CdT including particularly Maria-José Palos Caravina and John Beaven, as well as all members of the IATE project Expert Group, Steering Group and Work Groups, including particularly Jean-Marie Dufrasne (Chairman of the Expert Group), Christian de Villers (IDA Project Officer for the IATE project), Agustín Jimenez (Coordinator of the Data Structure Work Group), and Dimitri Theologitis (Coordinator of the Workflow Work Group), and last but not least the participants from the contractors Q&R and CST, including particularly Fotis Zografos and Bente Maegaard, for their invaluable written and verbal contributions to the work described in this paper.

References

Center for Sprogteknologi, 2000 *IATE Project Data Structure Work Group Proposal*, Copenhagen, Denmark

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

Le Meur A. and Gradzka, A., 2000 IATE: Draft for a formal analysis of EUROTERMS, EUTERPRE, EURODICAUTOM and TIS, Rennes, France, 77p

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

Quality and Reliability S.A., 2000 *Project Initiation Document: IATE Project – Services for the Development of an Interactive Terminology Database System,* Athens, Greece

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

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

Schmitz K-D., 2000, MARTIF, *SALT and related work on standards*, talk given at CdT

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