From έ, ú and ç to é, ü and ß: introducing paneuropean multilingualism into the CELEX databases

P.E. ALEVANTIS

CEC, Belgium

J. MARIN-NAVARRO

CEC, Luxembourg

Abstract: This article describes the problems solved in the context of the introduction of Greek and special Latin characters in the CELEX databases. In the present article the following points are developed: Implementation of international standards for the coding of characters and communication (problems of terminals, networks, hosts, open architecture); Retrieval problems and collating sequence implementation for different languages inverted files); User-interface and man-machine dialogue in different languages (keyboards); Data entry system with different filters (proprietary character codification versus standardised codes).

Keywords: *Multilingualism, Multilingual databases, Languages, European Communities.*

1. Introduction

CELEX (Communitatis Europaeae LEX) is the inter-institutional computerised documentation system for European Community law. It is a database that contains acts of Community legislation and case law with their full text, as well as bibliographical data on preparatory acts and parliamentary questions. CELEX is dissemi nated by the Commission of the European Communities (online and through intermediary hosts) (Ref 1).

CELEX has been established as a multilingual system because the Treaties establishing the European Communities provide for nine official working languages. This multilingual aspect of CELEX is not a mere luxury but rather a matter of primary importance for the European citizen: Community law supersedes national law in many cases and that is the reason why CELEX, the main distribution channel for Community law, is considered a pillar of the united Europe of 1993.

CELEX exists already in French, English, German, Dutch, Italian, Danish, Greek and Spanish; the Portuguese version is under preparation.

Since their creation, and until recently (March 1990), the Latin versions of CELEX did not contain special Latin characters (e.g. accented French or German letters); the code used for these versions was the basic Latin alphabet (ISO 646). Special characters were transliterated as

Online Information 90
11–13 December 1990
London, England.

one character and some as two (e.g. é, è, ê = e, ä = ae, å = aa, æ = ae, ç = c, ö = oe, ß = ss etc.). Furthermore, texts were transposed into upper case ('appauvrissement')

2. The challenge

An explosion of demand for the system by users outside the European Institutions, due to the acceleration of the establishment of the Internal Market, and the progressive setting up of new language versions of CELEX, stressed the need for the introduction of special Latin characters. If these could be provided for in the system, the implementation of an office automation environment in the services of the Commission would benefit by the integration of CELEX's full text capability.

In addition, the creation of a Greek CELEX database in the 'poor' 7-bit environment was impossible. One could think of replacing lower case Latin letters with upper case Greek ones (as was done in some implementations in Greece, see Ref 2), or replacing the Greek by the Latin alphabet (transliteration). The adoption of these solutions would give rise to serious problems; in fact, legal texts written only in upper case letters are not recognized as binding in Greece. On the other hand, the Community's legal texts in Greek do contain words in Latin ('sui-generis' decisions, 'ad hoc' committees, 'ESPRIT' and 'RACE' programmes) that cannot be translated or transliterated into Greek and can in fact constitute useful search terms in the context of a Greek database. Last but not least, the accents can in no way be omitted from Greek texts. In Greek, accents mark the syllable that is stressed and can play a definitive role in interpreting the meaning of a word. Words may be written the same way but their meaning varies, depending on the syllable stressed (Refs 3, 4). Of course that is also the case with other Latin languages such as French, German or Spanish.

All these considerations led to the conclusion that extended character sets were needed for the CELEX databases. Initially, the problem had to be solved for the Greek version of the base: what was needed was an 8-bit standard for the Greek-Latin alphabet. The standard chosen was the international standard ISO 8859/7, because that is what is required by legislation in force (Ref 5) and because it is identical to the corresponding Greek national standard (a serious concern for an information provider, like CELEX). At a latter stage, and on the basis of the experience gained by the setting up of the Greek version, special Latin characters were introduced in the Latin versions of CELEX. For these versions the ISO 8859/1 standard was chosen.

Consequently, all the elements of the computer environment had to be adapted, i.e. the standards had to be implemented in real world products, particularly in:

(1) terminals and printers;
(2) the Database Management System (DBMS);
(3) the programs that 'feed' data into the base;
(4) the networks through which users access the database.

The Commission's Informatics environment made the challenge even greater by stipulating conformity with its well-established Informatics Architecture (Ref 6). The policy of implementing international standards to facilitate intercommunication of computer systems is quite logical. The problem is that the market is not always willing to follow when the need is felt, so alternative options have to be considered in order to make a valid choice.

This has been particularly true in the case of CELEX. Dealing with a fully multilingual environment (full Latin plus Greek) can only be successful if two-byte codes are used. However, there are no low-cost terminals (or emulators) for such a codification, while existing DBMS software can hardly cope with 8-bit bytes, still less with 16-bit ones. Standardisation in that area is still immature (Ref 7), and will remain so should it become evident that it is not absolutely necessary for the codification of oriental languages (Ref 8).

But a fully multilingual environment was not truly necessary in the case of CELEX. It was well proven that Greek words could not constitute search terms in a full Latin environment, in the same way that words with special Latin characters could not be used for searches in a Greek-Latin environment.

On the other hand, a 7-bit environment with ISO 2022 extensions could only complicate matters. Special Latin characters are provided for in each 7-bit national standard; however, the use of the French standard would mean that the German or the Danish special characters would be ignored. The adoption of proprietary codifications was of course unthinkable as it would represent a clear violation of Community legislation in force (Ref 5). For all these reasons the choice of the ISO 8859 series of standards was the most efficient and realistic way to meet the challenge (Ref 9). Or at least to attempt to meet it.

3. *Meeting the challenge*

The creation of the Greek version of CELEX was the first step towards the solution of the problem (Ref. 10).

After choosing the ISO 8859/7 standard (that the Commission helped to adopt), the following points were addressed:

3.1 Terminals/printers

VT 220 character terminals were adapted to the standards (ISO 8859/7 for Greek and 8859/1 for Latin characters). This was of the utmost importance as it was the first step of the whole project. Fast printers were also adapted.

At the same time different emulators were evaluated from a multilingual point of view. In fact, existing PCs follow different internal codifications in different countries. In the PCs sold in each European country (or group of countries) the corresponding national characters are of course present. However, the IBM Code Pages used in these PCs do not conform with the ISO 8859 series. That is why, in order to access the Greek or the special Latin characters in CELEX, emulators are needed on which users can easily activate a translation of incoming characters.

3.2 Database Management System

Fortunately, the development team did not have to create a new DBMS specifically adapted for Greek. As the Commission is practising an active policy of enforcement of international standards in the context of its calls for tenders, the DBMS provider for the existing CELEX databases (in that case BULL S.A.) was obliged to adapt its product (MISTRAL) to the international standards. This covered various aspects, such as:

- the possibility of handling 8-bit coded data;
- the possibility of creating 8-bit coded inverted files with the appropriate collating sequence;
- the possibility of declaring 8-bit coded field names;
- the possibility of communicating with the system through 8-bit coded commands.

The last two prerequisites were especially necessary for the Greek version. In the case of the commands, it became evident from the beginning that, for full-text searching, if the commands were kept in Latin characters, the user would be obliged to key in at least 30% more information. This is due to the fact that Greek-Latin keyboards (as well as Russian-Latin or others) produce Latin and Greek in alternate levels, accessible through special keys or combinations of keys. In that respect it is not to be expected that the Common Command Language (ISO 8777) will be very ergonomic in such environments (see also Ref 2).

The final implementation provides also for the activation (by the user) of a filter that transposes all special characters emitted by the system into upper case . In that way an upward compatibility is ensured for users not willing to adapt their PCs to receive special Latin characters. These users can continue querying the bases as they have always been doing.

3.3 Data entry programs and procedures

Most of the CELEX data originate with the Office for Official Publications of the European Communities (OPOCE) on magnetic tape. The programs introducing these data into the CELEX databases have been adapted consequently as they contained variable parameters associated with each language version.

However, some of the textual data are introduced locally. For these data, special filters were developed to transcode the multiple byte (proprietary) coding of the multilingual word processing system (Q-one) used by Commission services, into ISO 8859 codes and vice-versa.

The same set of programs implemented in the context of the word processing system itself serve for downloading Greek-Latin or full Latin texts from the CELEX databases into Q-one documents. Of course, the implementation of standards in the terminals, the DBMS and the word processing system provided internal users of the system (i.e. officials of the European Institutions) with a unique workstation.

On the other hand, the controlled vocabulary part of CELEX is produced through the use of an automatic translation system that is based on multilingual tables. This part of the system posed the least of problems because it was already functioning in an 8-bit mode.

3.4 Networks

Introducing Greek and special Latin characters in CELEX would be useless if the users could not see them. For that to happen the access to the host system should be in an 8-bit, transparent mode. A special 8-bit port was set up to facilitate this access. At the same time, the programs handling the access to the bases have been adapted so that they could send the appropriate escape sequences to configure the multilingual terminals. Thus, access of both the Greek and the full Latin databases is possible through the same port and with the same terminal.

Communication however, gave rise to another problem. Most European networks permit the transfer of 8-bit codes in an X.28 context, but the messages emitted by the PAD are in 7-bit even parity. In practice, this means that a user with a terminal configured in 8-bit transparent mode necessary to access CELEX, will receive incomprehensible messages from the network. It is hoped that this situation will be dealt with in the future.

Conclusion

CELEX could not use the argument 'græca sunt non leguntur' for Greek texts. By introducing the Greek script in the databases and in fact in the Commission's Informatics Architecture, CELEX made possible the treatment of the special characters of other European languages as well. Should the CELEX team be asked to produce another Latin version (e.g. for an Eastern European language) or even a Cyrillic one (on the basis of other parts of the ISO 8859 series) its task should not be too difficult. Its choices would be based on the same principles. The first question one would have to answer is 'for how many users is access to be arranged for all the available languages at the same time and through the same terminal ?'. In the case that a fully multilingual environment is necessary, it may prove that Europeans need 16-bit characters much more than the Chinese or the Japanese.

On the other hand, CELEX is planning to introduce a preprocessor to analyse the full text morphologically, before introduction into the bases and to help the user during the queries. However, for such a module to be implemented, the modernisation plan for the whole system must be completed first.

Acknowledgements

We would like to express our thanks to the following persons who helped us to attain the objectives set:

Mr Walter de Backer, Director of Informatics, Commission of the European Communities, for making available the resources necessary for this project.

Ms Victoria Bensch, DBM CELEX, Commission of the European Communities, for her generous support throughout the project.

Mr Alain Leloup, programmer, S.I.E. Brussels, for adapting the programs of the CELEX application.

Ms Monique Vincent, Q-one support, Quadratron Brussels, and Ms. Sylvie Tarizzo, MISTRAL support, BULL Luxembourg, for helping with the adaptations of their respective products.

Address for correspondence:
Panagiotis E. Alevantis
CEC, IMCO 4/20A 200
rue de la Loi
B-1049 Brussels, Belgium.
Tel: +32 (2) 235.00.94
Fax: +32 (2) 235.70.12

References

1. For information on CELEX, please contact: Commission of the European Communities, EUROBASES Service, 200 rue de la Loi, ARL 3/2, B-1049 Brussels. Tel. +32 (2) 235.00.01, 235.00.03
2. Skourlas, C. '(Greek) National Documentation Center: Software Development for Public Databases Management', *Proceedings of Online Information '88*, London, Learned Information, 1988. Alevizos, T. *et al.* 'An information retrieval system for the National Documentation Center (in Greek)', *Proceedings of the 2nd Panhellenic Information Technology Conference*, Salonica, Greek Computer Society, 1988.
3. Alevizos, T. *et al.* 'Information retrieval and Greek-Latin text', *Proceedings of Online Information '88*, London, Learned Information, 1988.
4. Passias, A. 'Problèmes liés à l'élaboration d'instruments linguistiques du projet INNOMOS: Synonymie et polys dans le langage juridique hellénique', *2o Convegno Informatica Giuridica al servizio del Paese*, Roma giugno 1978.
5. 'Council Decision of 22 December 1986, on standardisation in the field of information technology and telecommunications (87/95/EEC)', *Official Journal of the European Communities*, Series L, No 36, 7/2/87, p. 31.
6. *Guidelines for an Informatics Architecture*, 4th edition, Luxembourg, Office for Official Publications of the European Communities (OPOCE), 1990 (ISBN 92-826-0275-3)
7. Work of ISO/IEC JTC 1/SC 2 on DP ISO 10646, Multiple octet coded character set.
8. Qiao, J. *et al.* 'Six-digit Coding Method', *Communications of the ACM*, **33**, 5, May 1990, pp. 491–494.

9. D'Cruz, M. *et al.* 'Character Sets of today and tomorrow', *Computer Standards and Interfaces*, **8**, pp. 199–208, 1988/89.

10. Alevantis, P. 'Creating the Greek CELEX database, technical or managerial challenge?' *Terminologie et Traduction*, No 1, pp. 11–21, Luxembourg, OPOCE, 1988.

International standards

ISO 646: Information Processing — ISO 7-bit coded character set for information interchange.

ISO 8859: Information Processing — 8-bit single-byte coded graphic character sets. Part 1: Latin alphabet No. 1; Part 7: Latin/Greek alphabet

ISO 2022: Information Processing — ISO 7-bit and 8-bit coded character set — Code extension techniques.

ISO 8777: Documentation — Commands for interactive text searching.