Article

Alice in the Wonderland of SGML: streamlining text entry in the CELEX databases

J. Marin-Navarro, P.E. Alevantis

CEC, JMO C2/25 Bâtiment Jean Monnet, Plateau de Kirchberg, L-2920 Luxembourg

Abstract: *This article describes the system used for the introduction of textual data into the CELEX full-text document databases. The solution implemented is based on the establishment of a text production database for the management and validation of texts before introducing them into the CELEX dissemination databases, and the management of structured documents described with the help of an SGML syntax.*

1. Introduction

The explosive evolution of office automation systems provided users with very powerful tools for the production and manipulation of documents. The concept of the 'document' itself evolved into a multimedia entity (by incorporating text, image and sound), while new ways of describing the structure of documents were developed (i.e. Office Document Architecture — ODA, Standard Generalised Markup Language — SGML). Today we refer to structured documents all too often.

On the other hand the evolution, if any, in the world of information retrieval has not been so spectacular. Commercially-available software for the management of document databases, especially large ones, is based on the same concepts as ten years ago. However, with the integration of information retrieval functions into office automation systems, one can expect document packages capable of storing and retrieving structured documents to become available in the very near future.

Today, many (if not all) documents published are produced in an electronic format and some of them, for instance legal texts, are loaded onto databases. Although these texts are available on magnetic media, the format used is a more or less publishing-oriented one; furthermore, the organiastion of the dissemination databases into which these texts are to be introduced is not oriented towards the easy updating of textual information. This is the reason why texts originating in publishing-oriented environments should be passed through production-preparation systems, independent of the dissemination systems into which they are to be loaded.

In this paper we will present the architecture of an already decentralised operational production system, used for loading textual data into dissemination databases operating in a multilingual environment. The system constitutes an application of the concept of structured documents based on the SGML standard (ISO 8879).

2. Context

CELEX (Communitatis Europææ LEX) is the computerised documentation system for European community law. CELEX is produced as an interinstitutional system (Commission of the European Communities, Council of Ministers, European Parliament, Court of Justice of the European Communities, Economic and Social Committee, Court of Auditors) and is made available to officials of Community institutions as well as to the public. (For information on CELEX, please contact: Commission of the European Communities, Service EUROBASES, 200 rue de la Loi, B-1049 Brussels. Tel: +32 235 00 01, +32 235 00 03.) CELEX is a multilingual system and is available in French, English, German, Dutch, Italian, Danish, Greek and Spanish; the Portuguese version is under preparation.

Technically speaking, the most important features of CELEX in its current form are those given below:

1. One dissemination document full-text database is available for each official language. As mentioned above, eight language versions exist (with a ninth to be added soon) with more than 120 000 documents each. The total space occupied is 1000 Mb per database.
2. Production and dissemination are performed on the same mainframe installed at the Commission's Computer Centre in Luxembourg.
3. Dissemination databases run under a full-text document DBMS.
4. Each CELEX document is made up of two parts:
   (a) a bibliographic (or analytical) part covering keywords (i.e. controlled vocabulary such as authors' names, subject matter, etc.), dates, classification codes;
   (b) the full text of the officially published documents.
5. CELEX documents are produced as follows:
   (a) *bibliographic* — a single master file is used which contains the analyses of the documents in a coded,

language-independent form, where the different language versions are automatically generated with the help of translation tables;
(b) *textual* — for each language version, the texts are loaded directly from magnetic tapes; special programs are used that treat each format separately.

Since the system went into operation in the mid-seventies, its structure has not evolved in a radical way. Consequently, management and maintenance have become progressively more expensive and complex. In order to address this situation the Commission decided to launch a project aimed at the modernisation of the CELEX databases. The two major objectives of this project are:

(1) to improve the user-friendliness of the access to the information in CELEX;
(2) to rationalise and improve the management of the system by adapting it to the Informatics Architecture of the Commission (OPOCE 1990) and integrating it with the office automation environment of the organisation. The data entry system was named ALICE (ALImentation CElex).

To attain the objectives set, the management decided to split the project into several sub-projects. The first of them was to provide the procedure for introducing textual data and was designated TEXTERFACE (TEXT intERFACE).

3. Architecture of the ALICE-TEXTERFACE system

The main function of ALICE-TEXTERFACE is to process texts from different sources in order to produce the textual part of CELEX documents. The system accepts different formats as input and generates files, grouped by language version, ready to be introduced into the dissemination databases. ALICE-TEXTERFACE is implemented on a local Unix mini-computer, while the dissemination databases run on a powerful central mainframe.

The most important of the input sources (corresponding to 80% of the database coverage) is the *Official Journal of the European Communities* (*OJ*) produced by the Office for Official Publications of the European Communities (OPOCE). The files used for the publication of the *OJ* are transformed from their original FORMEX format to an SGML-based format (FORMEX-SGML). The FORMEX format (Guittet 1984) was defined by the OPOCE in order:

> "to provide a detailed and structured method for recording information about the OPOCE's publications in computer-readable bibliographic record, for exchange purposes between two or more computer-based systems".

FORMEX attempts to unify two different approaches to the interchange of textual data, namely the CCF (Common Communication Format — based on ISO 2709) (UNESCO 1984) and SGML (ISO 8879).

Other accepted SGML-based formats include CJ-SGML (for data originating from the Court of Justice), EP-SGML (for data from the European Parliament), as well as the internal format CLX-SGML which is used for storing documents in the ALICE system.

These formats use different character sets. FORMEX — and consequently FORMEX-SGML — format uses a character set based on ISO 6937 with an extension mechanism in line with ISO 2022. CJ-SGML and EP-SGML use a proprietary non-ambiguous character set (EBCDIL) which is a multilingual variant of EBCDIC.

Figure 1 shows the architecture of the system.

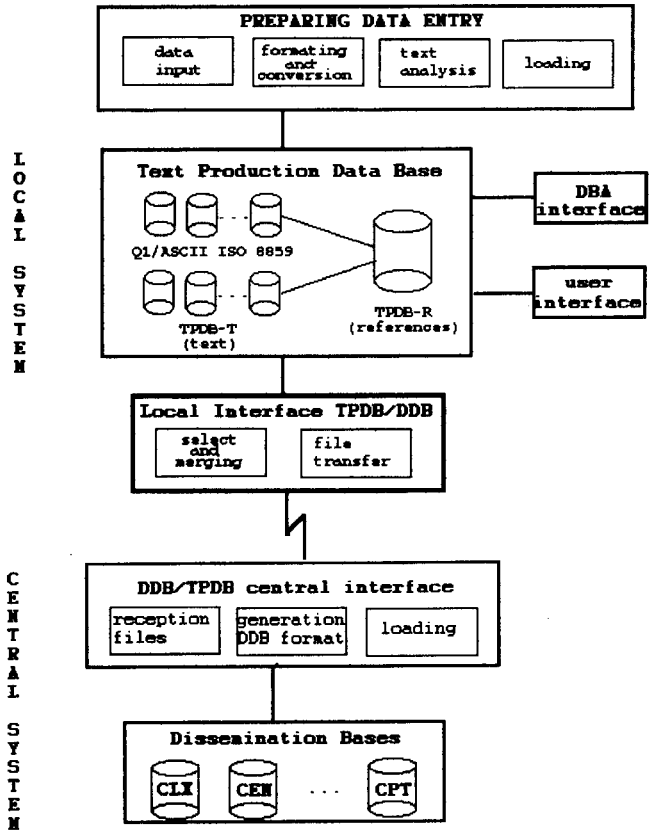

Figure 1: *Architecture of the ALICE-TEXTERFACE system.*

The most important modules are explained below.

3. 1. Text production database (TPDB)

The TPDB is the heart of the system. Its main purpose is the management and temporary storage of texts in the different language versions; these texts are to be validated, verified and eventually modified before they are definitively introduced into the CELEX dissemination databases.

The TPDB is structured as follows:

(1) a relational database containing the references (TPDB-R) of texts as well as statistical and management information;

(2) a set of sequential files (TPDB-T) containing the texts themselves, which are organised in a tree structure on the basis of the language of each document.

The link between the structured (TPDB-R) and the textual files (TPDB-T) is achieved by storing the name of the file containing the text of the document in a field of the TPDB-R.

3. 2. Text preparation

This module performs several functions.

1. Read/import of documents from different sources (CJ, EP, *OJ*).
2. Conversion into a unique character set (ISO 8859/1 for Latin-scripted languages, ISO 8859/7 for Greek) with transliteration of Greek characters found within Latin texts.
3. Validation and analysis of the structure of input documents in order to perform a logical division of the text appropriate to CELEX. The quality of the division depends on the structure of the original document and for some classes of documents the system cannot propose a standard structure. In this case, manual division is provided for. Finally, a set of files is generated: a file per published document plus a control file containing the references of each document.
4. During validation and analysis, the system also extracts information of an 'analytical' nature that will serve to initialise the system for the production of analytical data (ALICE-ARCHIVE) — this information is also structured with the help of SGML markers.
5. The resulting divided documents are stored in the TPDB-T and the TPDB-R is updated using the control file that is generated.
6. An SGML parser is used to validate the structure of input documents, and to perform the logical divisions during implementation of this module.

3. 3. User interface

Users (i.e. members of the CELEX management team) are equipped with a menu-oriented interface containing the usual functions associated with the management of a database: creation, display, modification, validation, deletion. Through this interface it is possible to access the reference files (TPDB-R) as well as the textual files (TPDB-T) of a document.

Access to the textual files is possible either via a simple screen editor or through a sophisticated word-processing package. Conversions between the character sets used in each case are provided for automatically.

Users can also introduce documents directly into the system. In this case, they have to introduce the markers themselves, in line with the CLX-SGML format. In the future it will be possible to replace the screen editor and the word-processing package with an SGML syntax-oriented editor that will make the online validation of the structure of documents possible.

All documents stored in the database remain there until they have been validated. Validation consists of the verification of the structure and the assignment of a CELEX number, the key to the whole system and the link with the analytical files.

3. 4. DBA interface

This module provides the specific functions linked with the management of the system (backup, restore, elimination of documents already loaded into the dissemination databases, etc.).

3. 5. Interface between text production and dissemination databases

Documents stored in the TPDB and already validated must be transferred to the dissemination databases. The module that performs this function is made up of two parts.

1. *Local*, which has the following functions:
   (a) selection of documents to be transferred;
   (b) validation and merging of selected documents as well as production of an SGML-based standard transfer format (a separate file is produced for each language version);
   (c) transfer of files to the central machine.
2. *Central*, which has the following functions:
   (a) reception of transferred files;
   (b) generation of the appropriate format for entry into the dissemination databases;
   (c) loading of the dissemination databases.

3. 6. Management of consistency between local and central databases

The system automatically checks for consistency between the text production database and the dissemination databases. This check must ensure that all documents are accounted for, while providing data for the clean-up at local level of documents already loaded onto the dissemination databases. The mechanism set-up uses dates linked with each stage of the production process of a document, as well as a process that consolidates the results of updating the dissemination databases in the local system.

4. Discussion of the solution adopted

The solution that has been implemented uses several concepts and techniques that should be discussed further.

4.1. Use of a production database for the management of texts

This choice is seen as practical and justified as a management tool for the following reasons.

1. Several data sources are used, with a need to validate and/or manipulate documents before their introduction into the dissemination databases.
2. The ability to access a document before its transfer to the central system greatly improves the quality of the dissemination database (textual corrections are easier to perform).
3. The time-span between the availability of published texts and their introduction into the dissemination databases can be dramatically reduced.

4.2. Separation of references (stored in a relational production database) and texts

This choice was not simply a means of overcoming the lack of powerful packages that can treat long texts in a Unix environment; it was a sound technical choice for several reasons:

1. It is not necessary to perform queries of a documentary nature while producing or manipulating a text, which means that functions offered by word-processing systems are sufficient.
2. It is necessary to construct an open system capable of adapting to evolving document concepts and allowing linguistic treatments (i.e. morphological analysis).
3. It avoids the restrictions and limitations of document DBMSs.

4.3. Use of SGML

Without making a check-list of the advantages and disadvantages of SGML, it must be emphasized that its most interesting feature is its simplicity and the ease of definition of document types. Thus, SGML is very well adapted to the exchange of information in a documentary context.

The use of SGML makes the definition of unrestricted exchange formats very easy. In the context of these formats, different logical elements of information are identified by markers. This is extremely interesting as far as the evolution of the application is concerned (new markers can be created without modifying the existing ones) while the difficulties associated with the treatment of other fixed formats are avoided altogether.

In the ALICE system, the SGML standard is extensively used for the definition not only of input formats (FORMEX-SGML, CJ-SGML, EP-SGML) but also for the basic format under which the documents are stored (CLX-SGML), as well as for exchange between the production system and the dissemination databases.

The use of an SGML parser is certainly an important advantage and facilitates the development process because it enables the format to be extended without problems. However, the absence of such a parser by no means constitutes a limitation with regard to the application of the SGML concept.

4.4. Open architecture independent of the DBMS used for the dissemination databases

The design and development of ALICE were guided by the need to implement a system capable of running on every Unix platform (POSIX in general) and eliminate, or at least minimise, restriction to any specific technology. The critical point of the CELEX system has always been the DBMS used for the dissemination databases. With the system in place and the definition of a basic format that is application-dependent and not DBMS-dependent, the generation of input files for other, different, packages is easily achievable and will not compromise the structure of the data input procedures.

5. Implementation and problems encountered

ALICE-TEXTERFACE was developed and runs on an NCR Tower 850 Unix mini-computer. The text production database was implemented under the ORACLE relational DBMS, while the textual files are accessible through the VI editor or the Q-one word-processing package. The dissemination databases run under MISTRAL, a document full-text DBMS, on a DPS 90 Honeywell-Bull mainframe under the GCOS 8 operating system. The SGML parser used was the MARK-IT package from SEMA GROUP S.A.

The first version of ALICE-TEXTERFACE was installed in September 1990, and the second is now available. The system permitted the introduction of all 1988 and 1989 texts that were missing from CELEX. It proved to be extremely user-friendly and flexible to use. Using the system it was possible to treat more than one *OJ* daily (all nine language versions). However, in order to accelerate the introduction of missing texts, it was decided to introduce only limited corrections in texts; because of this, the full capabilities of the system for correcting texts were not utilised.

It should be noted that the intrinsic complexity of the system, especially that of FORMEX, as well as the fact that some of the error messages of the MARK-IT parser were not quite explicit, made the detection of problems a cumbersome process.

More specifically, the major problems encountered can be summarised as follows.

1. Marking of *OJ* texts originating with OPOCE is not always accurate, since the texts are composed by different publishing houses, using different teams in different locations, which makes coordination difficult.
2. FORMEX SGML marking is presentation-oriented (markers are used to define different 'blocks' of text), the final product to be introduced into the dissemination

databases being marked according to content (preamble, corpus, annexe, etc.) and the transformation between the two is not always feasible.

3. The absence of a specific paragraph separator in the FORMEX-SGML context can sometimes complicate matters as a paragraph separator must be incorporated into the texts introduced into the databases to permit full-text searching in the same paragraph; a phrase separator is much more difficult to define as phrases in community legislation and case law are difficult to isolate.
4. As far as the character sets used are concerned, FORMEX implements a variant of the ISO 6937 standard with ISO 2022 extension techniques; however, the implementation is done in the context of a 7-bit environment, which posed some problems as some escape sequences are (very rarely) missing and the result may be a Latin text transliterated into Greek or vice versa.

6. Conclusion

Applying the concepts of open systems in the development of a decentralised application is not always easy, especially in the world of full-text databases. An independent production system for text entry can be made 'open'easily, through the extensive use of SGML-based formats for data interchange between the different sources of data as well as between the production and dissemination systems. The ALICE system will be further developed in order to cover the production of bibliographic files, as well as the introduction of special treatments for full text (e.g. morphological analysis).

Acknowledgements

We would like to express our thanks to the following persons and organisations who helped us attain the objectives set: Victoria Bensch, DBM CELEX and Project Manager, Commission of the European Communities, for her support throughout the whole project; Gilbert Joulain and Martine Renneson, programmers, Commission of the European Communities; Jean-Claude Xheunemont, SEMA Group Brussels, for developing and adapting the programs; and Emmanuel Albanese, Database Administrator of the ALICE-TEXTER-FACE system, CELEX team, Commission of the European Communities, for his valuable remarks on the use of the system and his intelligent implementation of the tools developed. Finally, we thank the Directorate of Informatics, Commission of the European Communities, for making the resources necessary for this project available.

References

GUITTET, C. (Ed.) (1984) *FORMEX, Formalized Exchange of Electronic Publications*, Luxembourg, Office for Official Publications of the European Communities (ISBN 92-825-5399-X).

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

ISO 2709 Documentation — Format for bibliographic information interchange on magnetic tape.

ISO 6937 Coded character sets for text communications (several parts).

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

ISO 8879 Information processing systems. Text and office systems. Standard Generalised Mark-up Language (SGML).

OPOCE (1990) *Guidelines for an Informatics Architecture*, 4th edition, Luxembourg, Office for Official Publications of the European Communities (ISBN 92-826-0275-3).

UNESCO (1984) *CCF: The Common Communications Format*, Paris, UNESCO (PGI-84/WS/4).

The authors

José Marin-Navarro

José Marin-Navarro received his Masters degree in Physical Sciences (Informatics option) from the University of Madrid in 1975 and completed a post-graduate Diploma in Computer Science at the University of Nancy, France in 1978. He worked for five years as Systems Analyst responsible for the development of information systems with the local government of Valencia, Spain, and five years as Assistant Professor in the Department of Informatics, University of Valencia, Spain, before joining the Informatics Directorate, Applications Engineering Department, of the Commission of the European Communities in Luxembourg as a Systems Analyst in 1986. He has been responsible for projects in the field of document databases and helped to introduce the Greek script into CELEX, serving as 'Systems supplier' for CELEX's modernisation project.

Panagiotis Alevantis

Panagiotis 'Takis' Alevantis received his Diploma in Physics from the University of Patras, Greece, in 1977. After serving for two years in the Greek army, he joined Papyros S.A., a Greek publishing house, in 1979, as assistant Editor, Science and Technology, for the *Papyros-Larousse-Britannica*, a Greek encyclopedia. He also worked as Science and Technology Editor on *Epikaira*, a weekly news magazine, and on *4Wheels*, a monthly car magazine. He joined the Commission of the European Communities in 1984 as a translator. From 1986 to 1991 he worked with the CELEX team, where he managed the project for the creation of the Greek version of the CELEX database and helped draft the specification for the modernised CELEX system. In May 1991 he joined the Informatics Department of the Translation Service of the CEC, where he serves as Project Manager for the introduction of extended multilingualism in the computer systems used by CEC translators.