Jouko Riikonen, WSOY Information Systems

WSOY editorial system for dictionaries

ABSTRACT: The paper describes the methodology and technical implementation of a very advanced editorial system which is used to assist the dictionary making process of the leading publisher in Finland.

1. Background

1.1 New editorial system

Currently WSOY has over 60 dictionaries on the market. These dictionaries were produced with truly traditional methods. There were many reasons to modernize the dictionary production system. Modern effective computer technology could not be fully utilized, traditional editing methods required much manual work and were prone to errors, it was practically impossible to use new electronic publishing medias, etc.

A decision was made to develop a new editorial system with modern logistics and the work was started in 1989. The complete system analysis, design and implementation has been done in-house. The reason for this was that there were no off-the-shelf solutions to meet the special requirements of lexicography and the Finnish language. Close co-operation with some software developers was however necessary as some modifications were needed to accomplish successful software integration. The initial requirements were better quality, shorter work cycle, re-use feasibility and better control of the entire publishing process (hard to guess). New publishing medias and rapid development in hardware and software technology were also taken into account.

Currently there are about 20 dictionaries under production with the new methods. Experiments to test the complete work cycle have been successfully completed. As a result new types of dictionary products have been released onto the market, e.g. 6 dictionary products on diskettes and a multilingual CD-ROM.

1.2 Basic approaches

The system modelling and philosophy is built around SGML (ISO Standard 8879) which facilitates an object-oriented view of the information manipulation. Documents (e.g. dictionary articles) together with respective DTDs (document type definitions) can be considered to be objects which are manipulated with different methods i.e. software

tools. Different kinds of properties can be assigned to different elements, groups or entire articles/documents, e.g. automatic content check like restricted vocabulary, spelling and hyphenation rules depending on the language of the element, grouping of the articles, special formatting etc.

The result is that structure, content and format are separated from each other. At the same time the system is relatively independent of the software or the hardware used, or the final publishing medium. Within the "SGML-world", all information exchanged between different systems is flexible. This also means that the best possible tools can be used for any particular task whether it is dictionary editing, data searching and retrieval or page composition.

For every dictionary product it is necessary to define a specific DTD, particularly in the case of bilingual dictionaries. However, it is possible to use certain global structures and element groups which are common to a wide class of dictionaries. The same DTD is then used during the whole work cycle. In the DTD development the elements and structures should reflect all the different aspects that may arise when handling the data with various tools. The basic structure is, however, specified according to the lexicographical needs.

2. The work cycle

2.1. Structure analysis

The SGML formalizes the way in which the dictionary articles are prepared. In addition to the structure and coding, SGML will settle how dictionary data can be handled and transferred between different computer systems and programs. The first task in a new dictionary project is the structure analysis of the new dictionary (and possibly also the source and the reference material). During the structure analysis the DTD for the new dictionary will be developed. WSOY has adapted a model where the leaves (dictionary elements) are first extracted from representative samples before the trunk (structure) is built. This work is done in co-operation with authors, dictionary staff and software specialists and will usually require many iterations. It will take from a couple of weeks to several months to conclude the final DTD.

2.2. Data input

Eventually all the data has to be converted to SGML. Existing dictionary material on typesetting tapes is analyzed for structure and parsed to match the respective DTD. Parsers are developed with XGML Translator (to be upgraded to Omnimark) and its XTRAN programming language (Software Exoterica, Canada). Typically, manual corrections are necessary after the parsing and are handled with a structure sensitive editor, the Checkmark (also from Software Exoterica).

Outside authors are urged to use structure-sensitive editing tools e.g. Author/Editor (Softquad Inc., Canada). Using structured editors the authors are more free to concentrate on the essential, the content itself. This is, however, not always possible and so special house rules based on SGML minimization have been developed which enable

authors to use normal word processors. This method (not optical character recognition) is also used for existing dictionaries which are not available in usable electronic form. House rules include also standardized entity names for special characters as these are not available in the character sets of normal PCs.

For in-house editing several editing tools are utilized including the Gestorlex (Textware A/S, Denmark) and SGML-editor (Arbortext, USA). Which tool is chosen depends on the nature of the work and available hardware platforms.

When complete SGML-coded dictionaries are not available for revision, so-called dictionary templates will be offered to the authors. Templates contain the framework (headwords, idioms etc.) for the new dictionary. As some dictionary publishers around the world have already adopted SGML, it will be quite feasible to co-operate and exchange the dictionary material and provide the above mentioned templates.

2.3. The database system

SGML-coded dictionary data is stored in a structured text DBMS, The Officesmith (Canadian Technology and Marketing Group Ltd). The software combines the editing capabilities of a word processor with the storage, retrieval and reporting capabilities of an on-line DBMS. It also fulfils the special requirements of dictionaries and Finnish, e.g. alternate collating sequences, unlimited character sets with the use of SGML entity feature and morphological word indexing. All fields of a document can be indexed in various ways (keyword, stopword, value, morphology etc.). Structuring allows the retrieval of documents or parts of the documents across different DTDs in a single query by defining a keyword. As the DBMS also supports SGML inclusions, handling of such in-field elements as optional and alternative elements, proper names, references and links etc. is possible early in the publishing cycle. In addition to traditional archive and retrieval applications, the database is also used for various content checks.

Applications to support dictionary work might also include a termbank for the collection of new words and phrases, corpora and Finnish language headword lists for dictionaries of different sizes. It is possible that simple text retrieval software e.g. public domain Texas/FreeText or PAT (Open Text Software, Waterloo, Canada) could be used especially in case of corpora as these do not require a very strict and sophisticated structure and are mostly used just for retrieving background information.

2.4 Output system

The outputting of the dictionary data has to meet various requirements. The output format can be as varied as a terminal screen, final dictionary pages or the file format for an electronic dictionary. For this purpose we use SGML based FOSIs (Formatting Output Specification Instances) which define a link between the format and the SGML documents. This facilitates also a fully automated page composition. Authors or editors make only final corrections (especially hyphenation) with the publishing software, so no expertise in handling the DTP-system is required from them.

In case of very complicated dictionaries where short hand notations are widely used, the possibility of making content errors is very high. To assist authors to find errors, special proof pages are generated where for instance abbreviations are opened, tildes are

replaced with the headwords, typography and layout are adapted to make it easy to spot critical points etc.

Currently, we use FOSI to define a filter which converts the SGML document instances to a FrameMaker file format (Frame Technology, USA). In the future we expect DTP-software to fully integrate the SGML support.

3. Technical highlights

3.1 Hardware and system software

The system hardware consists of IBM RS/6000 server, Sun Sparcstation2, 3 X-terminals, 3 PS/2s , 6 Macintoshes and Postscript laserprinters, some of which are shared with other units. Computers are connected to the department Ethernet, which is part of the company WAN (wide area network). Unix systems are preferred so TCP/IP, NFS, X11 and other de facto standards are followed where applicable.

KEYWORDS: dictionary publishing system, SGML, databases, structure sensitive editors, publishing tools, dictionary production cycle