◊ ◊ ◊

# *The Euromath System – a structured XML editor and browser*

J. Chlebíková, J. Guričan, M. Nagy, I. Odrobina
Faculty of Mathematics, Physics and Informatics
Comenius University, Bratislava

abstract. The Euromath System is an XML WYSIWYG structured editor and browser with the possibility of TeX input/output. It was developed within the Euromath Project and funded through the SCIENCE programme of the European Commission. Originally, the core of the Euromath System was based on the commercial SGML structured editor Grif. At present, the Euromath System is in the final stage of re-implementation based upon the public domain structured editor Thot and XML. During the re-implementation process several principal differences between the basic features of Thot and the basic purposes of the Euromath System had to be resolved.

keywords: structured editing, TEX, XML

## Introduction

During the last decade we have seen a revolution in the field of processing of the electronic documents. The rapid growth of information technologies brought significant changes to the potential benefits of electronic documents. The current evolution and the research of electronic documents has several basic goals:

- The document can be used for multiple purposes with different applications. Once a document has been stored in electronic form, one should be able to derive multiple "products" from a single source. For example, various kinds of printed material can be produced, the document can be used on the WWW, the document can be searched by database applications, and some parts of the document can communicate with external applications.
- The document has to have a long life-time. It should be easily revised and usable in any stage of its lifetime.
- Documents should be easily interchangeable across different computer platforms and networks.

In line with the previous goals, the markup of documents was developed. For multipurpose documents it is necessary to use general standardized markup with emphasis

on the logical structure of the document. For this reason the idea of a DTD (Document Type Definition) and a syntax taken from GenCode and GML were formalized and SGML (Standard Generalized Markup Language – ISO Standard 8879:1986) was developed (see [4]). Undoubtedly the most popular DTD is HTML, the present language of the WWW.

SGML is a complex standard, and its use remained limited mainly to large companies and a few research institutes. But this is not true for XML (Extensive Markup Language) — the new language of the WWW. The most important difference between XML and HTML is that XML does not have a fixed set of elements and attributes.

## Structured Editors

¿From the user's point of view the most comfortable tool for editing XML documents is a structured editor. A structured editor can guide the user according to the logical structure of the edited document (without his exact knowledge of that structure). In particular, a structured editor can prevent the user from producing a document whose actual logical structure is not consistent with the intended logical structure.

Advanced structured editors can allow the user to deal with the whole logical parts of the document in several ways:

- ◇ move or copy complete logical parts of the document,
- ◇ change an element to an element of another type,
- ◇ create or delete some additional structure around an element
- ◇ and so on.

It is not assumed that the author is familiar with the logical structure of the document. The editor offers options to the user according to the logical structure of the document.

Some structured editors can display and simultaneously allow one to edit individual elements of the logical structure in separate windows, for example the bibliography. It is possible to search for references to a particular logical element, and the editor may facilitate searching for logical elements of a specified type, e.g. tables or figures.

WYSIWYG structured editors have become the most popular; in these the presentation of a document is configured separately for each available logical structure. This approach uses a similar philosophy as LaTeX classes and has several advantages for the user. The author of the document only has to take care of the content of the document — the layout is produced automatically. Also several different presentations can be defined for one logical structure. The editor takes care of updating the numbering of theorems, footnotes, cross-references, etc. according to the logical structure of the document.

At present there are a few freely available WYSIWYG structured editors. We mention Thot, Amaya and the new version of the Euromath System presented here.

*Thot* ([5]) is an open experimental authoring system developed by the Opera project in order to validate the concepts of a structured document. Thot uses three different internal languages S, P and T for the manipulation of the document, but unfortunately it stores documents in the binary PIV format. From an abstract point of view the

S language (for Structure) provides the same structural concepts as a DTD. The P language (for Presentation) of Thot provides presentation or style sheet support to facilitate WYSIWYG views of documents. The T language (for Translation) allows one to define export specifications for each element (or rules for 'Save As' formats).

*Amaya* ([3]) is the W3C test-bed browser and authoring tool for HTML documents developed on top of the Thot technology. Amaya also has support for MathML (Mathematical Markup Language) and CSS (Cascading Style Sheet).

In the following we introduce the new version of the *Euromath System* — an XML authoring tool and browser based on Thot. It was developed in the Euromath Support Center (Faculty of Mathematics, Physics and Informatics) in Bratislava.

## Euromath System

The primary purpose of the Euromath System is to create a homogeneous computer working environment for mathematicians based on a uniform data model. The first version of the Euromath System (1992) was developed within the Euromath Project led by the European Mathematical Trust. Now the Euromath System combines the advantages of the WYSIWYG approach, structured editing and standardized XML format.

Originally, the core of the Euromath system was based on the commercial SGML structured editor Grif. At present, the core of the Euromath System is in the final step of re-implementation based on XML and Thot. Thot, unlike Grif, is public domain software which is also available for more platforms (Linux, Unix). Due to the conceptual proximity of both editors, the re-implementation was possible.

It is important to say that during the re-implementation process several principal differences between the basic features of Thot and the basic purposes of the Euromath System had to be resolved.

(i) *There is no direct support of XML in Thot.* The problem was solved by a new tool named `DTD2SPT` that is capable of porting an arbitrary XML DTD to the Euromath System. According to the DTD and a feature file the tool generates three files in the internal languages of Thot S, P and T. The S-file describes the logical structure and follows directly from DTD. The P-file is an automatically generated standard non-WYSIWYG 'XML' presentation, in which the logical tree structure of the document is displayed together with the tags for logical elements (which are usually hidden in other presentations). The T-file is used for saving the document in XML format. The user can influence the automatically generated S, P and T-files via certain rules in the feature file. These generated files customize Thot in such a way that it provides comfortable editing for documents which follow the rules of the given DTD. More detailed information about this tool can be found in Subsection "The DTD2SPT tool for porting a DTD into the Euromath System".

(ii) *Thot uses the binary PIV format for saving documents.* Therefore, it was necessary to solve the problem of opening and saving XML documents. Due to the fact that Thot has the possibility to add one's own T-language for export of the document, the latter problem was solved almost directly. The problem was solved by adding a

mechanism for opening XML documents. It mainly involves the translation of the input XML document into an internal S-structure according the document's DTD. This is one of the most important features of the Euromath System. Some key issues of the realization of the programme are given in Subsection "Opening XML documents".

(iii) The third important change follows from the fact that Thot is an authoring system, but the Euromath System is also a WWW browser.

*The Euromath system as structured WYSIWYG XML editor*

The Euromath System offers the same basic editing functions as non-structured text editors (find-replace, operations with clipboard, ...), and a spelling checker for English, French and other languages. It also offers the possibility to easily incorporate graphics according to various standards. The Euromath System allows easy WYSIWYG addition of new characters with support for UNICODE or the entity mechanism.

The Euromath System enables WYSIWYG structured editing for those DTD's that are frequently used by mathematicians: EmArticle.dtd, EmLetter.dtd, EmSheet.dtd, EmSlide.dtd, and others. These DTD's correspond to LATEX classes. As was mentioned before, the Euromath System facilitates structured editing of documents according any new DTD. But this requires that one write a presentation of the new DTD in Thot's P-language. Moreover, the user can add more than one presentation, so that one eventually has several different presentations for one DTD.

Furthermore, the Euromath System enables to use the advantages of the modularity approach. The user can defined some parts of the document structure as a modul, for example a table, etc. It is comfortable to use modularity approach in the case, if the parts of the document structure are identical for some document classes. Euromath System contain the moduls for paragraphs, tables and mathematical expressions (WYSIWYG presentations, saving into LATEX, ...). From these moduls the WYSIWYG presentation of new documents can be created rather straightforwardly. If the document consists of modules, the Euromath System allows one to change the presentation for each module during editing.

After re-implementation to the structured editor Thot, the Euromath System offers the same properties of the structured editor as before (see [1] for overview). The most of them was mention in the Section "Structured Editor".

The Euromath System stores a document automatically in XML format according the corresponding DTD. In the T-language the user has the possibility to add a translation of a document to other structured formats (for example, transform the structure to another one) and to unstructured formats like TEX or HTML. As the system deals with a structured document, the translation to most formats is easy. But for a perfect translation to TEX the T-language lacks certain features (more conditions, external files, etc.).

The Euromath System offers export of documents according to the EmArticle and EmLetter DTDs to the standard LATEX classes article, letter and book. For a new document class the translation to LATEX can easily be built from the available modules such as paragraphs, tables, formulae, etc.

## Euromath Applications

The Euromath System comes with a programming interface that allows external actions to be attached to it. Euromath applications extend the possibilities of the Euromath System as a structured editor.

*Personal File System (PFS)*

The Personal File System is a front-end for the ZentrallBlatt Math database. The form for formulating a (database) query is part of the Euromath System and uses the internal Thot library.

PFS is technically based on three external programs — `pcmes`, `zbl2tex` and `l2s`. `pcmes` comes from ZentrallBlatt, `zbl2tex` and `l2s` are part of the Euromath System and were developed in the framework of the Euromath Project.

`pcmes` is a database engine which executes queries formulated by the user using the PFS form, and generates record sets or other results. These results are returned to the Euromath System using a special listener.

`zbl2tex` transforms a record set obtained from pcmes to an XML file containing the required bibliographic data. This file is opened as a new document using the EmArticle DTD.

Some parts of a database record (TI-title, AB-abstract and UT-keywords) can contain TEX expressions and therefore must be processed by l2s. The source part of the record, which contains bibliographic data, does not have a fixed structure. Especially, books and conference articles can be very complicated and differ from case to case. We use a few heuristic methods to transform this part to the XML structure specified by the EmArticle DTD. These heuristics successfully cover more than 95 %. To get the resulting XML file, `zbl2tex` parses the original output from `pcmes` together with some auxiliary data in three stages.

*Translation from TEX to XML* – `l2s`

`l2s`, written for the Euromath project, is a parametrizable translation program, which translates TEX, LATEX and AMS-TEX files to files in XML format following the rules of EmArticle DTD (MathML DTD is in progress). Both the lexer and the parser parts contain some new features in order to cope e.g. with default parameters in LATEX2e macros, with the `ensuremath` construction or with `proclaim` in the amsppt style. l2s was mainly written for the translation of LATEX(2e) article style formatted documents to XML format.

The main part of `l2s` is also used for another way of inputting mathematics. The user can insert mathematics using either the standard means of a structured editor or by inserting TEX expressions as special elements. The translation between the TEX formulae and the XML structure can be done at any time by pressing `Ctrl-T` (or using the menu). Even when writing a text, one can press `Ctrl-T` to obtain the possibility to enter TEX formulae directly. When done, pressing `Ctrl-T` again displays the formulae in WYSIWYG mode.

The user can use his own predefined set of TEX macros given in a file determined by the MFILEL2S environment variable.

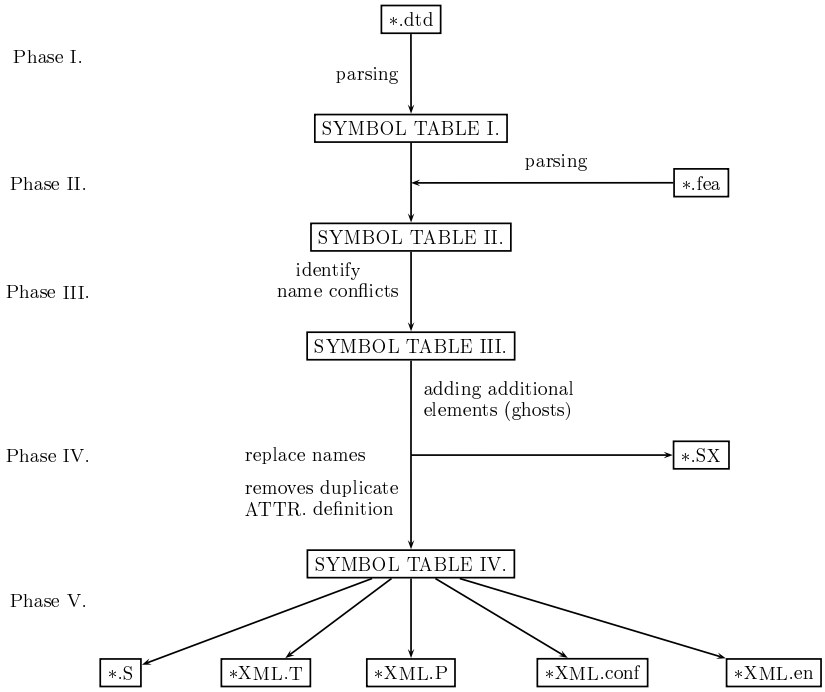


FIGURE 1: THE PHASES AND ACTIONS OF THE DTD2SPT. THE SX FILE IS AN AUXILIARY FILE DESCRIBING THE NAME CONVERSIONS OF THE OBJECTS.

*Retrieving documents across networks*
The Euromath System can be used as a WWW browser for HTML and XML documents. The Euromath System is ideal for viewing remote XML files – especially for documents with DTD's, for which WYSIWYG presentations in the Euromath System are available. The EmArticle DTD also contains some basic mathematical constructions but in the future we would like to add a WYSIWYG presentation for standard MathML.

## IMPLEMENTATION OF THE PROGRAMME

*The DTD2SPT tool for porting a DTD into the Euromath System*
The tool translates a DTD to files in the S, P, T languages, which are accepted by Thot. The translation process, inputs, outputs and internal states are shown in Figure 1. The tool passes through five phases, which are separated by four defined states of the principal internal data structure called the SYMBOL TABLE.

During phase I `DTD2SPT` reads the DTD and transforms it into its internal data structure. The user might find it useful to influence the automatic generation of the S, P, T language files by slightly changing the original DTD. This is enabled via

supplemental commands in the *feature file*. As can be seen in Figure 1, the DTD2SPT alters the internal representation of the DTD. This process is carried out in phase II and the SYMBOL TABLE is transformed to its second state.

During phase III DTD2SPT identifies those names of DTD objects (elements, attributes, attribute-definition-parts) which cannot be used in the S, P, T languages. This happens because the DTD specification permits names with a larger number of characters from a broader set; the DTD-object names may also interfere with keywords of the S, P, T languages.

In addition to these syntactical limitations, we had to deal with a dissimilarity in the background model of objects in the DTD and in the S, P, T languages. An attribute in the DTD is a sub-part of a single element, and the attribute-definition-part is a sub-part of an attribute. On the other hand, an attribute in the S-language is visible to all elements, and an attribute-definition-part to all attributes. This clearly implies that, unlike in DTDs, the names of all S-objects must be unique. Due to this dissimilarity, a straightforward translation of a DTD might also generate several definitions of a uniquely named attribute. Therefore during phase IV, attribute names and attribute definitions are compared; conflicts in definitions are resolved and redundant definitions are removed.

The content model of a DTD element can contain also group of elements or occurrence indicators. Here we encounter another problem that prevents a straightforward translation because generation of the presentation (the P-language) for this type of elements represents a serious complication. The problem is resolved during the same translation phase when these multi-composed elements are disassembled into elements with a simple structure. This process introduces additional elements (called ghosts) into the generated S-code.

*Opening XML documents*

The XML parser is the basic component of the Euromath System. The chosen approach uses the advantages of RTN (recursive transition network). The idea was adapted from [2], who had used RTN to parse natural language sentences. Allen's design was changed in detail to simplify the implementation.

The context-free grammar described in [6] was changed to RTN diagrams. Mainly, for each grammar rule one RTN was designed. The transition arc is labeled either by the terminal string or by a nonterminal one, which represents another RTN. An RTN is implemented as one function, which returns `accept`, `reject` or `error message`. When an RTN returns reject, it must restore the position of the read head. Problems could occur in the case of inclusions (internal or external entities) because it is time consuming to restore the original state before calling the RTN function. But the grammar rules in [6] are well proposed and the grammar is unambiguous. The unambiguity ensures that an RTN can decide whether to return reject or to continue processing (after which it can return only accept or error) by reading a few terminals from the tape.

Semantic rules have been implemented into the RTN functions. For example, the same name of the start-tag and the end-tag, the inclusion of entities, management of processing instructions, ... The inner structure of the document is built while the

input document is read with the assistance of Thot in accordance with RTN parsing. The validity of the document is checked automatically by the Thot engine according the internal structure of the S-language.

In order to successfully open XML documents, it was necessary to make some changes to Thot's S-language. For example, we introduced string conversion, new types of elements and so on.

Unicode has been a serious problem and it has not yet been fully resolved. The Thot engine does supports not it at all. In the present version, the large unicode numbers may be used as well a character entity references. Predefined named unicode entities can be automatically included into every document.

Changed presentation features of the document (for example, the font name or the font size) and the name of the actual presentation are stored as processing instructions.

## Concluding Remarks

XML is rapidly becoming the standard for the WWW. The Euromath System is at the forefront in exploiting the benefits of XML for documents and also the typesetting qualities of the TEX system.

The latest version of the Euromath system is available for UNIX (X-window system) on the SUN platform and Linux.

More information about the Euromath system can be found on the page `http://www.dcs.fmph.uniba.sk/~emt`.

Our further plans involve support of MathML, improving the possibility of structural changes and supporting at least some feature of CSS or XSL.

## REFERENCES

[1] J. Chlebíková, The Euromath System – the structured editor for mathematicians, EuroTeX 1998, pp. 82–93.

[2] J. Allen, Natural Language Understanding, 2nd ed., The Benjamin/Cummings Publishing Company, Inc., CA, 1995.

[3] V. Quint, I. Vatton, An Introduction to Amaya, World Wide Web Journal, Vol. 2, Num. 2, 39–46, 1997.

[4] International Standard ISO 8879, Information Processing – Text and Office Systems – Standard Generalized Markup Language, International Standard Organization, 1986.

[5] Opéra, Thot, A structured document editor, Inria, 1997. `http://www.inrialpes.fr/opera/Thot.en.html`

[6] XML specification, version 1.0, see `http://www.w3c.org/TR/REC-xml`.