Producing SGML-documents with public domain tools[1]

Paula Leinonen and Martti Penttonen

This report describes a project where a study guide was transformed to a structured form. Transformation was done, because the document was to be used both in printed form and in electronical form. Structure was marked up with SGML, and the structured document was converted to delivery formats with SGML-Tools. With some changes to the default system, the document was transformed to delivery formats, LaTeX and HTML, nearly automatically.

1 Introduction

Increasing telecommunication between organizations and new forms of documents demand new properties from text processing. Documents are handled in different environments. Marking up the logical structure of a document allows for reusing documents and transforming them to several delivery formats.

This report describes a project where a big, public document was transformed to a structured form. The document is the study guide of the Faculty of Science at the University of Joensuu, containing degree requirements for university students. The document is published yearly or every second year and the structure of it is quite fixed. The text content of the document alters slightly in every reprint. The book has about 200 pages and it has several writers, because the work process divides into many departments. The structure of the document was marked up with SGML (Standard Generalized Markup Language). SGML was chosen as the markup language for the following reasons. The document is produced in co-operation between six departments so SGML is a common standard, which was hoped to increase consistency of the document. The document is also structured by its nature and it is used as a reference book. The main purpose of the project was to produce a paper document, but the chance to get it also in HTML form at the same time was tempting. Multiple editing was hoped to be avoided.

No money was allocated for purchasing commercial software. An installation of Framemaker+SGML at the computer science department was available in the formatting phase of the project. In addition to this some freely available software, as SGML-Tools, were known to be available.

The secretaries at the six departments of the faculty, who were to type their chapters, had no experience about SGML or structured documents. Typists had PC's with DOS and Windows, but they had no SGML-based tools available. So they needed at least some kind of preview program for the on-line validation of the SGML document. The final, combined version of the study guide was made in UNIX. We had no earlier experience about this kind of projects or SGML tools for accomplishing paper document and HTML document from document with SGML tags.

[1]This document was done using SGML-Tools

2 Why SGML

Most electronic publishing systems used today are WYSIWYG-systems (What You See Is What You Get). These systems are quite easy to use for typists, but they do not support the structured approach. This kind of system usually uses some form of procedural markup which is understood only by this system in some particular environment. Procedural markup is often limited to the format and layout, and does not characterize hierarchy and element identification. Procedural markup codes are good for one presentation of the information, for example some printed page format. If the layout style changes or information is needed in a different format, the whole document must be reprocessed. If an organization changes software or hardware, translation to another coding format can be very difficult, time-consuming and expensive. If large documents have several authors, the result of combining parts can be incoherent without a common, controlling structure for directing the writing process and a common definition file for creating layout.

SGML is an international standard for describing the structure of the document. Descriptive markup describes the text as consisting of logical parts and makes possible to manage information as data objects rather than as characters on a page. Instead of describing printed output of the document, descriptive markup describes the syntactic role of the information in a document. The aim is to make the format of the document independent of any hardware and software environments and to separate the structure and the layout. Structured approach implies many benefits, like the organization of the information in a document, reusability and longevity of information. This kind of markup allows for outputting the information in several layouts and forms.

In SGML, *Document Type Definition* (DTD) describes the structure elements of the document and rules for defining relations between elements. Elements in a document can be title, chapters and sections for example. Style of the document can be defined separately, so writers can concentrate at the contents of the document rather than thinking about its layout. The structure helps authors in organizing the information. SGML is an open system and it can be used to represent multiple data formats like multimedia information, programming code or SGML queries. SGML is used to define the structure of the document but it does not standardize style or any other processing method. Some standards have been developed for defining the style of the SGML document, for example Formatting Output Specification Instance (FOSI) and Document Style Semantics and Specification Language (DSSSL) [5].

There were also drawbacks in using SGML. Changing from an old WYSIWYG system to SGML required more work than just updating the old document. Many typists of the document were unsatisfied because no WYSIWYG editor was available and they were not familiar with SGML before. The advantages of the SGML were not obvious for the users, who only encoded the document with SGML tags and it was difficult to motivate them.

3 How all was done

Because of scarce experience about SGML, the project was a trial-and-error process. We had to transform a Word Perfect document to SGML under DOS or Windows, we needed a SGML-previewer under DOS or Windows, and had to transform SGML to a paper document and to HTML in UNIX. We had several ideas about how to do the necessary sub-tasks in our project. First of all we gathered knowledge and software to be used in different parts of the project. Most of the time was used for finding and testing suitable tools for all phases including conversions from

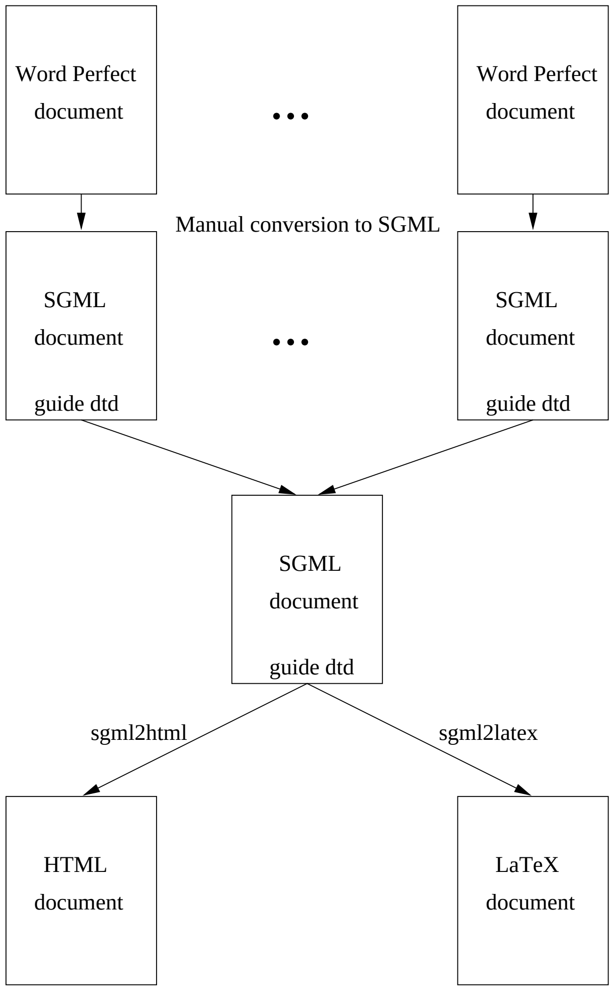


Figure 1: The document in the different phases of the project.

an unstructured document to SGML, from SGML to paper, and from SGML to on-line documents. SGML-Tools formatting system [4] was chosen for formatting of the document, because it is free and it allows for producing a variety of output formats. We decided to use it for checking SGML syntax and creating HTML output. At the beginning of the project it was not clear, what tool we would use for creating the paper document. Potential tools were `sgml2latex` provided by SGML-Tools and FrameMaker+SGML.

Figure 1 depicts the phases of the project as a graph. In Appendices A to **??** there are examples of the forms of the document during the project.

3.1 Structure of the document

Linuxdoc DTD provided by SGML-Tools was chosen as the basis of the document type definition of the study guide. It has been modified from QWERTZ DTD [1] for using particularly for Linux HOWTOs and other Linux documentation. QWERTZ DTD was originally a DTD for LaTeX. A few elements and an attribute was added to the DTD. The new DTD was named Guide DTD. The grammar in Figure 2, a subset of Guide DTD, was used as a guideline of coders.

```
article  ->  titlepag, toc?, p*, sect*, (appendix, sect+)?

titlepag ->  title, author
title    ->  pcdata | f | em | bf | sf | tt, (subtitle?) +(newline)
author   ->  name, thanks?, inst?, (and, name, thanks?, inst?)*
name     ->  pcdata| f | sq | em | bf | sf | tt
thanks   ->  pcdata| f | sq | em | bf | sf | tt
inst     ->  pcdata| f | sq | em | bf | sf | tt +(newline)
and      ->  empty

toc      ->  empty

sect     ->  heading, p*, sect1*
sect1    ->  heading, p*, sect2*
sect2    ->  heading, p*, sect3*
sect3    ->  heading, p*, sect4*
sect4    ->  heading, p*

heading  ->  (pcdata | f | em | bf | sf | tt )*

p        ->  (pcdata | f | em | bf | sf| tt )* | itemize | enum |
descrip |
              quote | tscreen | eng | small | tabular | table | verb |
              code )+) +(newline)

tabular  ->  (hline?, ((pcdata | f | sq | em | bf | sf| tt )*,
              (colsep, (pcdata | f | sq | em | bf | sf| tt )*)*),
               (rowsep, hline?, ((pcdata | f | sq | em | bf | sf| tt )*,
                (colsep, (pcdata | f | sq | em | bf | sf| tt )*)*))*)
hline    ->  empty
colsep   ->  empty
rowsep   ->  empty

quote    ->  (((pcdata | f | sq | em | bf | sf| tt )* | itemize | enum |
                         quote | tscreen | figure | tabular | verb)*, p*)+

small    ->  (((pcdata | f | sq | em | bf | sf| tt )* | itemize | enum |
                         quote | tscreen | figure | tabular | verb)*, p*)+

sq       ->  pcdata| f | sq | em | bf | sf | tt

itemize  ->  (item+)
enum     ->  (item+)
descrip  ->  (tag?, p+)+

item     ->  ((pcdata | f | sq | em | bf | sf| tt )* | itemize | enum |
                          quote | tscreen | figure | tabular | verb )*, p*
tag      ->  pcdata| f | sq | em | bf | sf | tt

tscreen  ->  (((pcdata | f | sq | em | bf | sf| tt )* | itemize | enum |
                         quote | tscreen | figure | tabular | verb )*, p*)+

eng      ->  (pcdata | f | sq | em | bf | sf| tt )*

verb     ->  rcdata
code     ->  rcdata

em       ->  (pcdata | f | sq | em | bf | sf| tt )*
bf       ->  (pcdata | f | sq | em | bf | sf| tt )*
sf       ->  (pcdata | f | sq | em | bf | sf| tt )*
tt       ->  (pcdata | f | sq | em | bf | sf| tt )*

appendix ->  empty

empty    ->  pcdata
```

Figure 2: The structure of the study guide

3.2 Conversion of source text to SGML

Since the study guide has been published yearly, the main part of the text existed. The existing document was written with Word Perfect text processing system. At the beginning of the project the typists had to make necessary changes to the text content. Conversion to SGML was done manually by clerks in the five departments of the faculty. The Word Perfect -document was saved as a plain text document and SGML tags were added to the document with Word Perfect or some other text editor. Instructions about how to write documents for SGML-Tools was provided to the typists. The instructions were based on illustrative examples about marking all text elements needed in the document. It would have been possible to make some automatic preliminary coding, but we did not find tools for doing it in time.

Two different environments, DOS and UNIX, caused the next problem to the project. The parser of the SGML-Tools was available for DOS, but for the conversion we had to find other tools. We needed conversion to plain text, HTML or LaTeX, because typists needed some kind of preview-program when they were adding SGML tags to the document. At first, before finding a conversion program, we planned to send the SGML document from PC to UNIX machine through network and for conversion with SGML-Tools. We however found tools for doing conversion in PC and a preview-program was composed of several freely available programs. These programs were combined to one application with a graphical user interface by Visual Basic (Fig. 3). See Appendix D for further information about the programs and the ftp sites, where the programs used in the project are available. For *normalizing* minimized markup, `sgmlnorm` in James Clark's SP package was used. It was necessary because `sgml2tex`, which was needed later, does not support markup minimization. The normalized document instance was converted to LaTeX by `sgml2tex` program, which converts the SGML tags in a document to TeX control sequences. The `sgml2tex` program produces code output, which includes empty definitions for all elements, attributes and entities encountered, in a dummy style file. The user has to implement the style file in order to achieve the desired result. After it has been done, `sgml2tex` creates a LaTeX source code from the SGML file. EmTeX was used to make a PostScript file from the LaTeX code in DOS. LaTeX is a high quality typesetting system, which is based on TeX and EmTeX is a DOS implementation of TeX and LaTeX. The writers of the document could view the PostScript file by GSview PostScript previewer for OS/2 and MS-Windows. It took a lot of time to find and test all these programs and make them work together. The result is also quite massive. It requires about three megabytes disk space and it is slow. Its advantage is however, that the result of the conversion is almost identical to the final UNIX output when `sgml2latex` from SGML-Tools package is used.

In future, we are going to test if it is possible to use `emacs` -editor with `psgml` macro package for editing SGML source documents in the departments.

3.3 Conversions to delivery formats

Parts of the document were combined to one SGML instance, which was uploaded to a UNIX server. Two alternatives were considered for doing the paper version of the document, `sgml2latex` from SGML-Tools and FrameMaker+SGML. The first printing of the document was done using FrameMaker+SGML. Because the document contains a lot of data in the form of large tables, we thought that it would be easier to make the final layout of the document by using FrameMaker+SGML rather than by LaTeX. Description about using FrameMaker+SGML is in [2].

Figure 3: Previewer for the writers of the document.

Unfortunately, we could not automatize the conversion with FrameMaker+SGML satisfactorily. Difficulties followed from the different principle and the implementation of tables in these applications. QWERTZ DTD defines LaTeX-like tables and FrameMaker+SGML understands only CALS table structure, which is defined in [3]. Because widths of columns of tables were not marked in the markup phase of the project, they had to be defined manually with editing tools provided by FrameMaker+SGML. It was quite laborious and time-consuming. We had some alternatives to automatize the tranformation process. It should be possible to mark up the tables in detail by using CALS DTD, and then transform them automatically by using FrameMaker+SGML. It is however difficult to mark up tables so precisely and that is why we ultimately decided to test conversion with `sgml2latex` from SGML-Tools.

In section 3.3.2 we consider the conversion of the document by `sgml2latex`. We recommend to use this system in the next reprint of the document.

3.3.1 Conversions by SGML-Tools

SGML-Tools contains transformations to plain text, groff source, LaTeX, HTML, GNU Info, LyX and RTF. The document is first parsed and then transformed. Parsing is done with `sgmls` parser provided by James Clark. In this phase the mark-up of the document is verified so that it follows the structure defined by the DTD. The minimization of mark-up is normalized by expanding all macros and elements for the next phase of the process. An accurate and complete sgml document is transformed to another format with `sgmlasp` and a *replacement file* for the target format. In the replacement file, the correspondence of the elements in the SGML document and the elements in the target document is defined. Because linuxdoc DTD is a LaTeX-like DTD, the conversion from SGML to LaTeX is quite easy. For example element `<title>` is converted to LaTeX according to the next two lines:

```
<title>         +       "\\title{"
</title>                "}"             +
```

The program replaces the tag `<title>` in SGML file with `\\title{`, joins the title text to that string and ends it with } during the transformation process.

The user can expand the replacement file with other definitions or substitute file with new definitions. If the structure of the target document differs considerably from the linuxdoc DTD, it is more difficult to do the mapping. It is also possible to make an own special DTD and replacement file suitable for it.

3.3.2 Paper document

The document contains a lot of large tables. At first we were not aware how to do long tables with LaTeX and therefore used FrameMaker+SGML with CALS tables. After the paper document was already produced with FrameMaker+SGML, we found the `longtable`-package, which was selected for doing tables because it can handle tables which are longer than one page. A sample page of the final document is seen in Figure 4.

In addition to making transformation of the tables more automatic, we decided to fix some other weaknesses, met in the conversion from SGML to FrameMaker+SGML document. For example, the hyphenation of paragraphs with different languages was easily automatized.

For automatizing formatting of tables we decided to specify the markup of the tables with widths of the columns in some situations. Widths of the columns are marked up using marking `p{`*width*`}`. It handles the contents of the column like a paragraph and divides long texts to several lines.

For example, marking columns of the table like

```
<tabular ca="p{2cm}p{12cm}">
```

formats the table by dividing long columns to several lines:

8.15	Kansainvälisten asiain yksikkö esittäytyy, apulaisihteeri Pirkko Ojala
9.00	Oy Amica Ab -ravintola esittäytyy ja tarjoaa kahvit
9.45	Ylioppilaiden terveydenhoitosäätiön palvelut, lääkäri Eija Vänskä ja terveydenhoitaja Seija Käyhkö
10.30	Opiskelijaseurakunta esittäytyy, opiskelijapappi Tiina Tarnanen
10.45	Joensuun yliopiston ylioppilaskunta esittäytyy, Ari Maskonen ja Jari Mäkinen

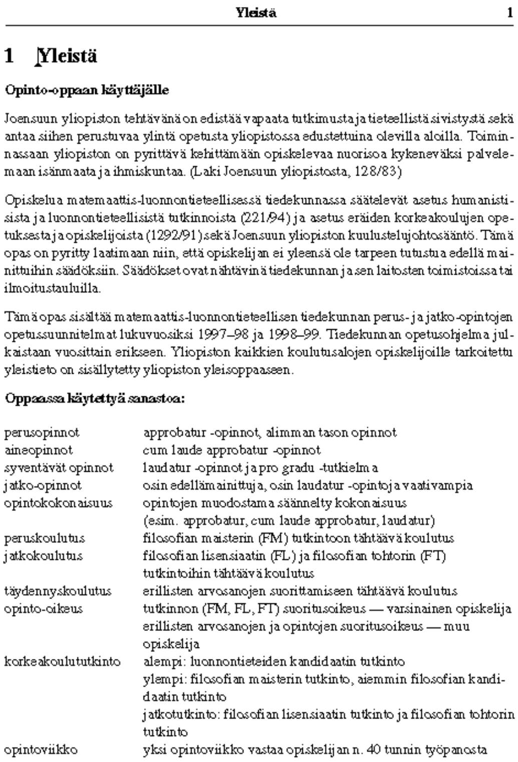
Part of the definitions of the widths of the columns was done automatically by defining a proper `sed` script. The rest was done manually.

For doing conversion of the tables from SGML to LaTeX nearly automatic, only two changes in the default system were necessary. First, widths of the columns were specified in some cases and then `tabular` environment of LaTeX was replaced with `longtable` in the replacement file.

The document is mostly in Finnish, but there are also some English paragraphs. For hyphenating the English paragraphs, the element `hyp` was added to the DTD. In transformation process tags were handled by the following lines in the replacement file:

```
<hyp>          "\n\\selectlanguage{[LAN]}\n"
</hyp>
```

After transformation, there is a correct markup in the LaTeX source file such that all paragraphs are hyphenated in a right way.



Yleistä 1

1 Yleistä

Opinto-oppaan käyttäjälle

Joensuun yliopiston tehtävänä on edistää vapaata tutkimusta ja tieteellistä sivistystä sekä antaa siihen perustuvaa ylintä opetusta yliopistossa edustettuina olevilla aloilla. Toiminnassaan yliopiston on pyrittävä kehittämään opiskelevaa nuorisoa kykeneväksi palvelemaan isänmaata ja ihmiskuntaa. (Laki Joensuun yliopistosta, 128/83)

Opiskelua matemaattis-luonnontieteellisessä tiedekunnassa säätelevät asetus humanistisista ja luonnontieteellisistä tutkinnoista (221/94) ja asetus eräiden korkeakoulujen opetuksesta ja opiskelijoista (1292/91) sekä Joensuun yliopiston kuulustelujohtosääntö. Tämä opas on pyritty laatimaan niin, että opiskelijan ei yleensä ole tarpeen tutustua edellä mainittuihin säädöksiin. Säädökset ovat nähtävinä tiedekunnan ja sen laitosten toimistoissa tai ilmoitustauluilla.

Tämä opas sisältää matemaattis-luonnontieteellisen tiedekunnan perus- ja jatko-opintojen opetussuunnitelmat lukuvuosiksi 1997–98 ja 1998–99. Tiedekunnan opetusohjelma julkaistaan vuosittain erikseen. Yliopiston kaikkien koulutusalojen opiskelijoille tarkoitettu yleistieto on sisällytetty yliopiston yleisoppaaseen.

Oppaassa käytettyä sanastoa:

perusopinnot	approbatur -opinnot, alimman tason opinnot
aineopinnot	cum laude approbatur -opinnot
syventävät opinnot	laudatur -opinnot ja pro gradu -tutkielma
jatko-opinnot	osin edellämainittuja, osin laudatur -opintoja vaativampia
opintokokonaisuus	opintojen muodostama säännelty kokonaisuus (esim. approbatur, cum laude approbatur, laudatur)
peruskoulutus	filosofian maisterin (FM) tutkintoon tähtäävä koulutus
jatkokoulutus	filosofian lisensiaatin (FL) ja filosofian tohtorin (FT) tutkintoihin tähtäävä koulutus
täydennyskoulutus	erillisten arvosanojen suorittamiseen tähtäävä koulutus
opinto-oikeus	tutkinnon (FM, FL, FT) suoritusoikeus — varsinainen opiskelija erillisten arvosanojen ja opintojen suoritusoikeus — muu opiskelija
korkeakoulututkinto	alempi: luonnontieteiden kandidaatin tutkinto ylempi: filosofian maisterin tutkinto, aiemmin filosofian kandidaatin tutkinto jatkotutkinto: filosofian lisensiaatin tutkinto ja filosofian tohtorin tutkinto
opintoviikko	yksi opintoviikko vastaa opiskelijan n. 40 tunnin työpanosta

Figure 4: The printed document.

The attribute `fsize` was added to the element `article` for providing a suitable fontsize in the SGML file.

```
<!element article - -
        (titlepag, header?, toc?, lof?, lot?, p*, sect*,
                          (appendix, sect+)?, biblio?) +(footnote)
<!attlist article
        opts cdata "null"
        fsize cdata "12pt">
```

In the replacement file the attribute `fsize` is handled by the line

```
<article>      + "\\documentclass\[[FSIZE]\]{article}\n"
```

which sets a value of the attribute to the variable FSIZE.

3.3.3 HTML document

The SGML version of the document with linuxdoc tables was used as a starting point for the conversion to a HTML document. Conversion to HTML was easier than conversion to a paper document. This is because there are no limitations of paper in HTML pages. For example, handling of long tables is not needed.

`Sgml2html` from SGML-tools package was used to convert the document. The default conversion of `sgml2html` can be changed or expanded. We used an extra replacement file for table tags for getting columns, because the default replacement file does not define them. The expansion contains the following definitions:

```
<tabular>      +       "<TABLE><TR><TD>"            +
</tabular>     +       "</TD></TR></TABLE>"         +

<rowsep>               "</TD></TR><TR><TD>"         +

<colsep>               "</TD><TD>"
```

The rest was done automatically but some ruled tables with column spanning were finished manually after conversion.

Automatic conversion by `sgml2html` produced table

```
<TABLE>
<TR> <TD><HR></TD> <TD></TD> <TD></TD> </TR>
<TR>
    <TD></TD><TD>Kieli- ja metodiopinnot 13 ov</TD><TD></TD>
</TR>
<TR> <TD><HR></TD><TD></TD><TD></TD> </TR>
<TR>
    <TD>Maantieteen </TD>
    <TD>Biologian cum laude </TD>
    <TD>Aineenopettajan</TD></TR>
```

```
<TR><TD>cum laude approbatur </TD>
    <TD>approbatur 35 ov </TD>
    <TD>pedagogiset opinnot</TD>
</TR>
<TR><TD>35 ov </TD>
    <TD>Vapaasti valittavat </TD>
    <TD>35 ov</TD>
</TR>
<TR><TD></TD>
    <TD>opinnot 5 ov </TD>
    <TD></TD>
</TR>
<TR> <TD> <HR> </TD><TD></TD><TD></TD> </TR>
<TR>
    <TD></TD><TD>Maantieteen laudatur 37 ov</TD><TD></TD>
</TR>
<TR> <TD> <HR> </TD> </TR>
</TABLE>
```

This table could be manually cleaned to

```
<TABLE border=1 width="100%">
<TR>
   <TD colspan="3" align="center">
   Kieli- ja metodiopinnot 13 ov </TD>
</TR>
<TR>
    <TD> Maantieteen </TD>
    <TD> Biologian cum laude </TD>
    <TD> Aineenopettajan</TD></TR>
<TR><TD> cum laude approbatur </TD>
    <TD> approbatur 35 ov </TD>
    <TD> pedagogiset opinnot</TD>
</TR>
<TR><TD> 35 ov </TD>
    <TD> Vapaasti valittavat </TD>
    <TD> 35 ov</TD>
</TR>
<TR><TD></TD>
    <TD> opinnot 5 ov </TD>
    <TD></TD>
</TR>
<TR>
    <TD colspan="3" align="center">Maantieteen laudatur 37 ov</TD>
</TR>
</TABLE>
```

There are only three tables of this kind in the document, so the work is not too big, and not worth automatizing.

4 Conclusions

In the first trial of the project, as described in [2], the goals were mostly reached, but the project required more work than it was estimated. It took about four months from the beginning to the end. However the work was finished in schedule. The paper document was sent to the University press as a PostScript file in electronic form. At the same time the link to the HTLM version of the document was joined to the web page of the Faculty of Science. The biggest problem was the tables, that demanded too much manual work after automatic conversion.

By rethinking, direct transformation to LaTeX with SGML-Tools was used, as described in this report, and the result was more automatic. Some additions to the original linuxdoc DTD were done, and the replacement file was updated respectively. Changes to the default system are mostly connected to formatting of the document, not to the logical structure of it. However, after marking up these insertions to the SGML document once, the conversion to LaTeX can be done more automatically. There are only few tables in the document which have to be edited manually after the conversion. Editing of those tables takes less than one hour, if the user is familiar with LaTeX.

The most critical part of the system is now the markup phase of the project, even if the biggest work from text document to SGML document has been done. The coders have no SGML tools for marking up the document. Document is updated yearly and there are always some changes to its content. It would be easier for the typists to write document with some WYSIWYG system or even some SGML-related tool.

A Unstructured, plain text document

```
MATEMAATTIS-LUONNONTIETEELLISEN TIEDEKUNNAN OPINTO-OPAS
Lukuvuosina 1997-1998 ja 1998-1999

Yleistä

Opinto-oppaan käyttäjälle

Joensuun yliopiston tehtävänä on edistää vapaata tutkimusta ja tieteellistä
sivistystä sekä antaa siihen perustuvaa ylintä opetusta yliopistossa
edustettuina olevilla aloilla. Toiminnassaan yliopiston on pyrittävä
kehittämään opiskelevaa nuorisoa kykeneväksi palvelemaan isänmaata ja
ihmiskuntaa. (Laki Joensuun yliopistosta, 128/83)

Opiskelua matemaattis-luonnontieteellisessä tiedekunnassa säätelevät asetus
humanistisista ja luonnontieteellisistä tutkinnoista (221/94) ja asetus
eräiden korkeakoulujen opetuksesta ja opiskelijoista (1292/91) sekä
Joensuun yliopiston kuulustelujohtosääntö. Tämä opas on pyritty laatimaan
niin, että opiskelijan ei yleensä ole tarpeen tutustua edellä mainittuihin
säädöksiin. Säädökset ovat nähtävinä tiedekunnan ja sen laitosten
toimistoissa tai ilmoitustauluilla.

Tämä opas sisältää matemaattis-luonnontieteellisen tiedekunnan perus- ja
jatko-opintojen opetussuunnitelmat lukuvuosiksi 1997-98 ja 1998-99.
```

```
Tiedekunnan opetusohjelma julkaistaan vuosittain erikseen. Yliopiston
kaikkien koulutusalojen opiskelijoille tarkoitettu yleistieto on
sisällytetty yliopiston yleisoppaaseen.

Oppaassa käytettyä sanastoa:
perusopinnot        approbatur -opinnot, alimman tason opinnot
aineopinnot         cum laude approbatur -opinnot
syventävät opinnot  laudatur -opinnot ja pro gradu -tutkielma
jatko-opinnot       osin edellämainittuja, osin laudatur -opintoja vaativampia
opintokokonaisuus   opintojen muodostama säännelty kokonaisuus
                    (esim. approbatur, cum laude approbatur, laudatur)
peruskoulutus       filosofian maisterin (FM) tutkintoon tähtäävä koulutus
jatkokoulutus       filosofian lisensiaatin (FL) ja filosofian tohtorin (FT)
                    tutkintoihin tähtäävä koulutus
täydennyskoulutus   erillisten arvosanojen suorittamiseen tähtäävä koulutus
opinto-oikeus       tutkinnon  (FM, FL, FT) suoritusoikeus - varsinainen opiskelija
                    erillisten arvosanojen ja opintojen suoritusoikeus -
                    muu opiskelija
korkeakoulututkinto alempi: luonnontieteiden kandidaatin tutkinto
                    ylempi: filosofian maisterin tutkinto, aiemmin filosofian
                    kandidaatin tutkinto
                    jatkotutkinto: filosofian lisensiaatin tutkinto ja filosofian
                    tohtorin tutkinto
opintoviikko        yksi opintoviikko vastaa opiskelijan n. 40 tunnin työpanosta
```

B Structured document with guide DTD markup

```
<!doctype linuxdoc system>

<article>

<title>MATEMAATTIS-LUONNONTIETEELLISEN TIEDEKUNNAN OPINTO-OPAS
Lukuvuosina 1997-1998 ja 1998-1999
<author><newline>
<toc>
<sect>Yleistä
<p>
<bf/Opinto-oppaan käyttäjälle/

Joensuun yliopiston tehtävänä on edistää vapaata tutkimusta ja tieteellistä
sivistystä sekä antaa siihen perustuvaa ylintä opetusta yliopistossa
edustettuina olevilla aloilla. Toiminnassaan yliopiston on pyrittävä
kehittämään opiskelevaa nuorisoa kykeneväksi palvelemaan isänmaata ja
ihmiskuntaa. (Laki Joensuun yliopistosta, 128/83)

Opiskelua matemaattis-luonnontieteellisessä tiedekunnassa säätelevät asetus
humanistisista ja luonnontieteellisistä tutkinnoista (221/94) ja asetus
eräiden korkeakoulujen opetuksesta ja opiskelijoista (1292/91) sekä
Joensuun yliopiston kuulustelujohtosääntö. Tämä opas on pyritty laatimaan
niin, että opiskelijan ei yleensä ole tarpeen tutustua edellä mainittuihin
```

```
säädöksiin. Säädökset ovat nähtävinä tiedekunnan ja sen laitosten
toimistoissa tai ilmoitustauluilla.

Tämä opas sisältää matemaattis-luonnontieteellisen tiedekunnan perus- ja
jatko-opintojen opetussuunnitelmat lukuvuosiksi 1997-98 ja 1998-99.
Tiedekunnan opetusohjelma julkaistaan vuosittain erikseen. Yliopiston
kaikkien koulutusalojen opiskelijoille tarkoitettu yleistieto on
sisällytetty yliopiston yleisoppaaseen.

<bf/Oppaassa käytettyä sanastoa:/
<tabular ca="lp{10cm}">
perusopinnot       |approbatur -opinnot, alimman tason opinnot@
aineopinnot        |cum laude approbatur -opinnot@
syventävät opinnot |laudatur -opinnot ja pro gradu -tutkielma@
jatko-opinnot      |osin edellämainittuja, osin laudatur -opintoja vaativampia@
opintokokonaisuus  |opintojen muodostama säännelty kokonaisuus@
                   |(esim. approbatur, cum laude approbatur, laudatur)@
peruskoulutus      |filosofian maisterin (FM) tutkintoon tähtäävä koulutus@
jatkokoulutus      |filosofian lisensiaatin (FL) ja filosofian tohtorin (FT)@
                   |tutkintoihin tähtäävä koulutus@
täydennyskoulutus  |erillisten arvosanojen suorittamiseen tähtäävä koulutus@
opinto-oikeus      |tutkinnon  (FM, FL, FT) suoritusoikeus --- varsinainen opiskelija@
                   |erillisten arvosanojen ja opintojen suoritusoikeus ---
                    muu opiskelija@
korkeakoulututkinto|alempi: luonnontieteiden kandidaatin tutkinto@
                   |ylempi: filosofian maisterin tutkinto, aiemmin filosofian
                    kandidaatin tutkinto@
                   |jatkotutkinto: filosofian lisensiaatin tutkinto ja filosofian
                    tohtorin tutkinto@
opintoviikko       |yksi opintoviikko vastaa opiskelijan n. 40 tunnin työpanosta@
</tabular>
```

C Converted structured document with HTML markup

```
<HTML>
<HEAD>
<TITLE>MATEMAATTIS-LUONNONTIETEELLISEN TIEDEKUNNAN OPINTO-OPAS: Yleistä</TITLE>
</HEAD>
<BODY>
Edellinen
<A HREF="opas-2.html">Seuraava</A>
<A HREF="opas.html#toc1">Sisällysluettelo</A>
<HR>
<H2><A NAME="s1">1. Yleistä</A></H2>

<P><B>Opinto-oppaan käyttäjälle</B></P>
<P>Joensuun yliopiston tehtävänä on edistää vapaata tutkimusta ja tieteellistä
sivistystä sekä antaa siihen perustuvaa ylintä opetusta yliopistossa edustettuina
olevilla aloilla. Toiminnassaan yliopiston on pyrittävä kehittämään opiskelevaa
nuorisoa kykeneväksi palvelemaan isänmaata ja ihmiskuntaa. (Laki Joensuun yliopistosta,
```

```
128/83)</P>
<P>Opiskelua matemaattis-luonnontieteellisessä tiedekunnassa säätelevät asetus
humanistisista ja luonnontieteellisistä tutkinnoista (221/94) ja asetus eräiden
korkeakoulujen opetuksesta ja opiskelijoista (1292/91) sekä Joensuun yliopiston
kuulustelujohtosääntö. Tämä opas on pyritty laatimaan niin, että opiskelijan ei
yleensä ole tarpeen tutustua edellä mainittuihin säädöksiin. Säädökset ovat nähtävinä
tiedekunnan ja sen laitosten toimistoissa tai ilmoitustauluilla.</P>
<P>Tämä opas sisältää matemaattis-luonnontieteellisen tiedekunnan perus- ja
jatko-opintojen opetussuunnitelmat lukuvuosiksi 1997-98 ja 1998-99. Tiedekunnan
opetusohjelma julkaistaan vuosittain erikseen. Yliopiston kaikkien koulutusalojen
opiskelijoille tarkoitettu yleistieto on sisällytetty yliopiston yleisoppaaseen.</P>

<P><B>Oppaassa käytettyä sanastoa:</B>
<TABLE>
<TR><TD> perusopinnot </TD><TD>approbatur -opinnot, alimman tason opinnot
</TD></TR>
<TR><TD> aineopinnot </TD><TD>cum laude approbatur -opinnot</TD></TR>
<TR><TD> syventävät opinnot </TD><TD>laudatur -opinnot ja pro
gradu -tutkielma</TD></TR>
<TR><TD> jatko-opinnot </TD><TD>osin edellämainittuja, osin
laudatur -opintoja vaativampia</TD></TR>
<TR><TD> opintokokonaisuus </TD><TD>opintojen muodostama säännelty
kokonaisuus</TD></TR>
<TR><TD> </TD><TD>(esim. approbatur, cum laude approbatur, laudatur)
</TD></TR>
<TR><TD> peruskoulutus </TD><TD>filosofian maisterin (FM) tutkintoon
tähtäävä koulutus</TD></TR>
<TR><TD> jatkokoulutus </TD><TD>filosofian lisensiaatin (FL) ja filosofian
tohtorin (FT)</TD></TR>
<TR><TD> </TD><TD>tutkintoihin tähtäävä koulutus</TD></TR>
<TR><TD> täydennyskoulutus </TD><TD>erillisten arvosanojen suorittamiseen
tähtäävä koulutus</TD></TR>
<TR><TD> opinto-oikeus </TD><TD>tutkinnon (FM, FL, FT) suoritusoikeus -
varsinainen opiskelija</TD></TR>
<TR><TD> </TD><TD>erillisten arvosanojen ja opintojen suoritusoikeus - muu
opiskelija</TD></TR>
<TR><TD> korkeakoulututkinto </TD><TD>alempi: luonnontieteiden kandidaatin
tutkinto</TD></TR>
<TR><TD> </TD><TD>ylempi: filosofian maisterin tutkinto, aiemmin filosofian
kandidaatin tutkinto</TD></TR>
<TR><TD> </TD><TD>jatkotutkinto: filosofian lisensiaatin tutkinto ja
filosofian tohtorin tutkinto</TD></TR>
<TR><TD> opintoviikko </TD><TD>yksi opintoviikko vastaa opiskelijan n. 40
tunnin työpanosta</TD></TR>
<TR><TD> </TD></TR></TABLE>
</P>
```

D Free software used in the project

- *SGML-Tools* package for SGML parsing and formatting is available from `ftp://sunsite.unc.edu/pub/Linux/utils/text/sgml-tools-0.99.0.tar.gz`.

- *SP* package for SGML parsing and entity management can be found at `ftp://ftp.jclark.com/pub/sp/sp-1.1.1.tar.gz`
- `sgml2tex` for DOS converts SGML documents to TeX. It is available from `ftp://curia.ucc.ie/pub/sgml/sgml2tex.zip`
- EmTeX is TeX and LaTeX for DOS and is available from `ftp://ftp.dante.de/tex-archive/systems/msdos/emtex`
- GSview is PostScript previewer for OS/2 and MS-Windows. It is freely available from `ftp://ftp.cs.wisc.edu/ghost/rjl/gsview21.zip`
- Adapted version from `sgml2latex` of SGML-Tools package is implemented during the project described in this paper. Package contains also example of the document in the SGML form. It is available from `ftp://ftp.cs.joensuu.fi/pub/Software/study_guide2latex.zip`

References

[1] T. Gordon: *The qwertz SGML Document Types, Version 1.2 Reference Manual*, 1994.

[2] P. Leinonen, M. Penttonen: SGML-dokumentin teko julkisohjelmilla. *Proceedings of SGML Finland 1997*, (Ed. K. Rytkönen), SGML Users' Group Finland, 1997, 10–29.

[3] Markup Requirements and Generic Style Specification for Electronic Printed Output and Exchange of Text, MIL-M-28001B.

[4] M. Welsh (updated by G. Hankins): SGML-Tools User's Guide, 1996.

[5] B. Travis, D. Waldt: *The SGML Implementation Guide: a Blueprint for SGML Migration*, Springer-Verlag, 1995.