Some Aspects of Presentation and Conversion of Scientific Documents (Online/Print)

Burkhard Kirste

Institut für Organische Chemie, Freie Universität, D-14195 Berlin

Abstract. SGML based concepts for the presentation of scientific facts as well as several tools for converting documents are discussed. In particular, the treatment of mathematical formulas, chemical data (molecular structures) and bibliographic data is dealt with. Scientific texts with sufficient "markup" allow easy conversions for online presentations and professional printouts and should simplify the preparation of excerpts for use in data bases.

Introduction

In the age of electronic technical information, it is necessary to be able to present scientific documents efficiently in different form: online in the World Wide Web (Internet) or a corporate intranet, as a professional hardcopy in print and possibly in further variants. "Efficiently" means that the conversion process should be essentially automatical.

In this paper several possibilities of currently available methods shall be discussed. For some conversions scripts have been developed here.

In principle, the task requires a "master file" containing all factual information as well as the formatting information ("markup") directly or indirectly (via auxiliary files, "style sheets" etc.). The application of SGML (Standard Generalized Markup Language) would be state of the art. Several parsers (sgmls, nsgmls) and DTDs (Document Type Definitions) as well as converters are freely available, e.g., qwertz-DTD, linuxdoc, gf with snafu-DTDs. It stands to reason that it should be possible to incorporate documents which are available in one of the popular formats such as Rich Text Format (RTF, from Microsoft Word) or LaTeX, and conversely, these formats should also be available as output formats. A further aspect would concern, e.g., the import of literature references from the results of database searches.

The inherent limitations of the different representations of documents deserve attention. Thus, there is no hypertext functionality in printed articles as a matter of fact, whereas the application of special symbols in HTML (Hypertext Markup Language, the language of the World Wide Web) currently is faced with difficulties. Although the Portable Document Format (PDF, Adobe) is more powerful with respect to the representation of special symbols and layout than the current versions of HTML and also offers hypertext functionality, it is proprietary and by no means an adequate substitute for HTML.

Using SGML (Standard Generalized Markup Language)

Whereas ordinary word-processing and desktop publishing systems produce rather unstructured documents, SGML (Standard Generalized Markup Language) deals with *structured* documents[1]. Moreover, SGML has been adopted as an ISO standard (ISO 8879) in 1986. Since only the ASCII character set is used, there are no conversion problems between different platforms (e.g., DOS, Unix, VAX). Thus, there is no dependence on a particular company, and SGML documents will remain legible and useful in the future.

It should be emphasized that SGML per se is only a *language*. For use with real documents (*instances*), a specific document type definition (DTD) is required. In short, a DTD defines the basic structure of the documents, the tags to be used and the (character) entities.

Some Freely Available SGML Systems

Here we will focus on SGML systems which are freely available. They are based on sgmls or its successor, nsgmls (SP), by James Clark (source code and binaries for various platforms are available).

First, the qwertz SGML document types by Thomas F. Gordon and the associated (Unix) tools shall be mentioned. The main purpose of the qwertz system is the representation of mathematics and the conversion to LaTeX documents. However, document types for letters, manual pages and bibliographies are also given, as well as tools for conversion to nroff (hence ISO latin-1 or ASCII text), grops and man pages.

Whereas conversion to HTML (HyperText Markup Language) is not covered by the original qwertz system, this is offered by the Linuxdoc-SGML formatting system by Matt Welsh and Greg Hankins which is based on the qwertz system. As the name implies, the Linuxdoc system was developed to produce documentation for the Linux operating system (Linux HOWTO's). However, it can be used as a general purpose system; the supplied replacement mappings may have to edited, particularly with respect to mathematics. The Linuxdoc system (version 1.5) offers the following output formats: HTML, LaTeX (2.09 or LaTeX2$_\varepsilon$, also DVI and PostScript), LyX, RTF (Rich Text Format[1]) and plain text (ASCII and ISO-8859-1, via groff). Although it is a bit of a "hack" and not perfect, it is quite useful. There is a simple example of a Linuxdoc document with SGML source and different output formats.

The third and last system which shall be mentioned is the general formatter (gf) by Gary Houston with the snafu SGML document type. Although gf is most powerful with snafu DTD SGML input, it is a particularly useful tool for converting HTML documents to LaTeX (LaTeX2$_\varepsilon$), plain text (ASCII or ISO-Latin-1) or RTF.

[1] RTF may be imported into Microsoft Word or compiled to a Windows help file. Currently, the RTF output of linuxdoc is prepared for use with the Windows help compiler.

Converters to and from HTML and SGML

It is fairly easy to write SGML documents directly with any text editor, particularly convenient are text editors with macro capabilities. This task is comparable to writing LaTeX or HTML documents. However, it is important to obey to the document structure; SGML syntax checkers are available. WYSIWYG SGML systems are available commercially but shall not be considered here.

It stands to reason that many people prefer a WYSIWYG word processor to an allegedly inconvenient text editor. Also, a vast amount of scientific literature exists already in electronic form, but not as (sufficiently) structured documents. (In the worst case, printed documents might be scanned and converted to ASCII texts by OCR.) Thus, converters capable of generating SGML documents from, say, LaTeX, RTF or HTML source will be helpful. (RTF, the Rich Text Format, is a document interchange format designed by Microsoft.) Since the structure that can be inferred from these sources is most likely insufficient, some manual formatting will usually be required.

Although HTML is based on SGML, it must be kept in mind that it was developed with the purpose of quick online formatting capabilities. Hence current versions of HTML are inferior to full-featured SGML DTDs, and the representation of complex scientific documents in HTML presents problems. Moreover, although the fairly primitive version HTML 2.0 (RFC 1866) is essentially a standard, current extensions are not (HTML 3.2, Netscape HTML Extensions, W3C HTML Experimental).

Many tools (filters) are available for converting other types of documents such as word processor output to HTML. Only two famous converters shall be mentioned here: the LaTeX2HTML translator (LaTeX to HTML) by Nikos Drakos and rtftohtml, a filter to translate RTF (and hence Microsoft Word documents) to HTML, by Chris Hector.

Current browsers (Netscape Navigator, Mosaic) can convert HTML documents to text or PostScript. The gf formatter has been mentioned above. Another tool for converting HTML to PostScript is html2ps by Jan Kärrman.[2]

Two (experimental) Perl scripts have been written here, latex2sgml.pl and html2sgml.pl, which may help in converting LaTeX or HTML documents to SGML documents (Linuxdoc DTD). These tools may have to be edited to suit particular needs, and some manual formatting of the resulting SGML document is generally necessary. Another tool which may be helpful in the process of converting documents is charconv; the purpose of this tool is the conversion between different extended character sets (e.g., DOS, Macintosh, ISO Latin 1, HTML, SGML), it does not provide any formatting.

Another approach which is worth mentioning is the MATHS system developed by Richard J. Botting. It provides a simple means of an ASCII representation of formal mathematics (BNF: Backus-Naur forms and the like) as well as tools for converting this notation (reusable mathematical information) to

[2] Mosaic (2.7) and html2ps allow pagination of the printout. However, the quality of the printouts is only moderate as compared to that achievable with LaTeX.

HTML. Two Perl scripts have been written here, mth2html and mth2sgml, which convert MATHS documents (allowing for several general-purpose extensions) to HTML and SGML (Linuxdoc DTD).

Figures Figures have to be converted as well. (Encapsulated) PostScript is commonly used for the preparation of printed documents, wheras GIF or JPEG are popular image formats for screen presentations; there are many other formats (e.g., TIFF, PBM/PPM, XBM, RGB, PCX, WMF, BMP), but converters are available (e.g., xv, imagemagick, netpbm/pbmplus, GhostScript, paint shop, wmf2bmp).

For instance, PostScript figures can be converted to GIF images by means of the pstogif script which is part of the LaTeX2HTML package; it makes use of GhostScript and netpbm/pbmplus. The conversion of GIF images to encapsulated PostScript can be achieved by netpbm/pbmplus (e.g, giftopnm followed by pnmtops) or by convert from the imagemagick package.

mdl2gif is a convenient tool for converting molecular structures (2D) given as MDL mol files to GIF images.[3]

Math Mode in HTML Documents

The qwertz SGML system (and hence Linuxdoc) offers a fairly good set of tags and character entities for the representation of mathematical formulas, although the coverage is not as comprehensive as that provided by TEX or LATEX. The snafu SGML system takes a different approach by using the TEX notation directly, embedded in <texeqn> tags, but that method is only useful for generating LATEX output.

Currently available WWW browsers do not support math mode, with the exception of the experimental Arena browser. (Tools like LaTeX2HTML convert mathematical formulas to graphics for use with ordinary browsers.) Mathematical formulas and the character entities of greek letters (e.g., α) have to be embedded in between <math> tags. It should be mentioned that the tool html2ps is able to convert (a subset of) math mode to PostScript. As an example, see a page dealing with a few formulas about the chemical shift. After the provided replacement mappings have been edited appropriately, the Linuxdoc system can convert SGML documents containing math mode to HTML output suitable for the Arena browser.

Extension to CML — Chemical Markup Language

In chemistry, the incorporation of two- and three-dimensional molecular structures into documents is of utmost importance. Currently structural formulas are supplied as graphics for print media. For hypermedia presentation, chemical

[3] mdl2gif, written by Roger Sayle, can be found in the CML package.

MIME (Multipurpose Internet Mail Extensions) may be used to supply coordinates and allow interactive online viewing of molecular models. As an extension of HTML, CML (Chemical Markup Language) has been suggested by Peter Murray-Rust, Henry S. Rzepa and Christopher Leach (see their poster) to include 2D and 3D structural information directly in SGML (CML) documents. Likewise, diagrams or (uuencoded or PostScript) figures may be included as well as numeric data in scalar or matrix form. Currently CML is in an experimental stage, there is a prototype viewer (cmlcost) for CML or ESIS files (End System-Intermediate System, generated by an SGML parser such as sgmls or nsgmls). CML documents are not supposed to be created manually but by means of tools.

Bibliographies

Online searches in bibliographic data bases yield bibliographic references in a format which is not directly suitable for use in manuscripts. Several commercial programs are available (e.g., Reference Manager or VCH Biblio) which allow import and export of bibliographic data. However, the BibTeX system and accompanying tools (e.g., bibview) offer an attractive and powerful alternative (for free).

A Perl script stn2biblio.pl has been written here for the conversion of bibliographic STN references (BIB format) to the SGML biblio DTD (which is part of the qwertz package), BibTeX, Unix refer/grefer or NLM format. (Unix refer with the lookbib/glookbib or lkbib tools offers a simple way of searching private bibliographic data bases.) The SGML biblio format in turn may be converted to BibTeX, refer, PostScript or ASCII text; by means of the script biblio2html.pl, HTML output can also be generated.

Conclusions

A more rational approach in dealing with scientific information is required. That means, all relevant documents should be available in an electronic format, perhaps SGML, that contains sufficient *markup* to allow for an automatic generation of the document formats requested by the user: high-quality print, online hypertext, excerpts for use in data bases etc.

The current waste of time (and money) caused by media breaks, manual reformatting, manual preparation of excerpts or even re-typing of text and redrawing of structural formulas must be avoided. Since many scientists prefer a WYSIWYG approach to directly generating SGML by editing a plain ASCII file, the corresponding tools will have to be developed or improved.

Online version of this paper, URL:
http://www.chemie.fu-berlin.de/chemistry/papers/cic96/

References

1. Brian E. Travis and Dale C. Waldt. *The SGML Implementation Guide. A Blueprint for SGML Migration*. Springer-Verlag, Berlin, 1996.