

XML-based text & graphics integration

Christian Märtin

Siemens AG, Information and Communication Networks, Central Organization, Munich, Germany
christian.maertin@icn.siemens.de

Jürgen Krüger

Siemens Business Services, Structured Document Processing, Munich, Germany
juergen.krueger@mch.sbs.de
<http://www.sbs.de>

Abstract:

Based on the Microsoft Office 2000 package and aligned with the current SGML editing system, the 'Office Solution' presented here shows the next evolutionary step in the application of SGML/XML and Web technologies. The development of the solution focused on intelligent, interactive graphics in combination with a proprietary navigation system and an effective model for linking text and graphics. The system is integrated into a higher level information and management system with defined access rights. ActiveX Control cannot be implemented at the present time due to security reasons.

Background

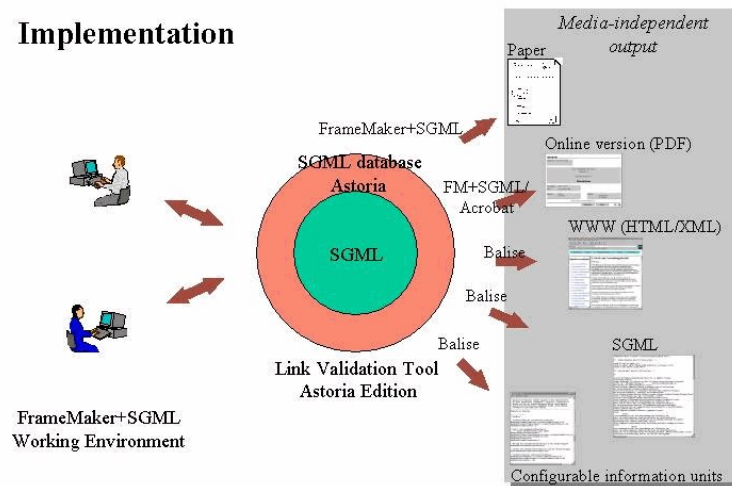


Figure 1. SGML editing system

In October 1998, an SGML-based editing system was introduced in the Information and Communication Networks Group at Siemens AG for the purpose of making the internal guidelines, work instructions and circulars available in the Intranet in HTML format.

The system concept and prototype were presented at SGML/XML Europe '98 in Paris .

The following aspects of the editing system used have remained open because of future extensions to HTML and XML functionality – including graphics options – by the W3C.

- Integration of active, intelligent graphics
- Introduction of an effective link model
- Migration to XML

New XML based system

During XML structural design, it was demonstrated that DTDs can be greatly simplified. Complicated link constructs originating in proprietary SGML data entry tools were completely redesigned. The premise on which the new structural design was based was that it uses few elements, attributes and content modeling in a simple main structure including graphical components. The structure should be applicable to many types of documents and it should be easy to edit the documents. More complex structures should apply these definitions to form the basic component.

The evaluation of different graphic formats and the task of finding a solution based on Microsoft Office 2000 yielded the following conclusions:

Microsoft Office 2000 should be used as a text and graphics editor.

Graphics should be integrated as an XML application – i.e., as VML. Microsoft Internet Explorer 5 can display VML without any plug-ins.

Links should be extracted from text and graphical documents and edited in a separate document.

These aspects are considered in the XML system design. By structuring and correcting the Office 2000 output with special transformation scripts (Balise), valid XML can be generated from text and well-formed XML from graphics.

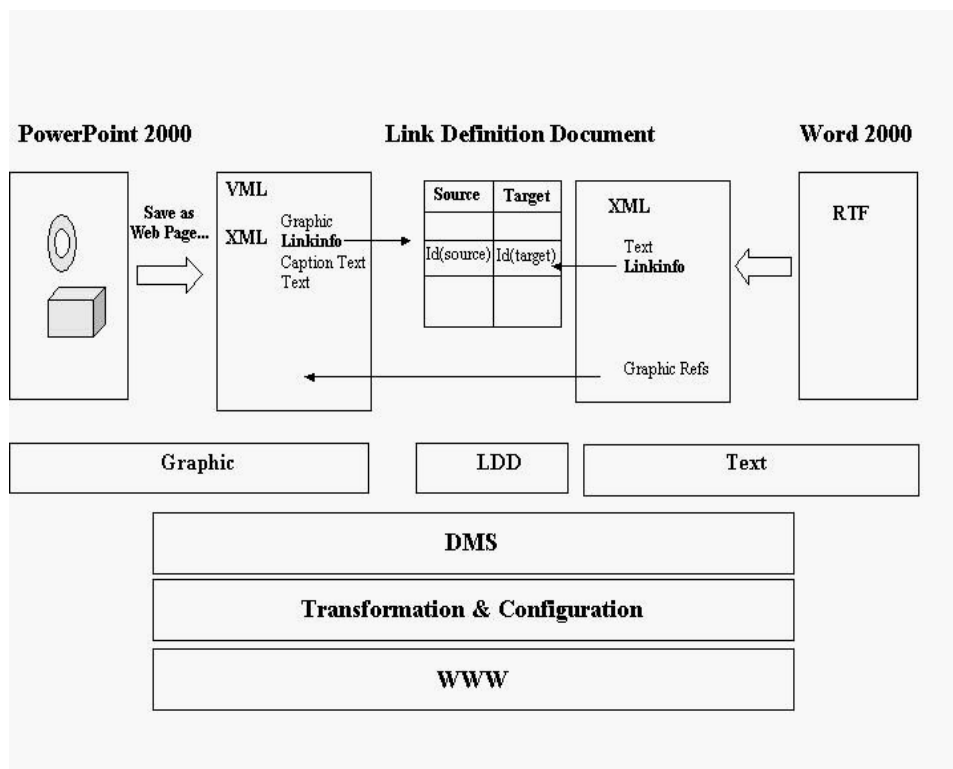


Figure 2. XML text/graphics system

Text component

Since Word 2000 is used as a typographical editor and not as a structure editor, text must be transformed into parsed, valid XML. This approach guarantees the automated editing functionality within a predefined structure.

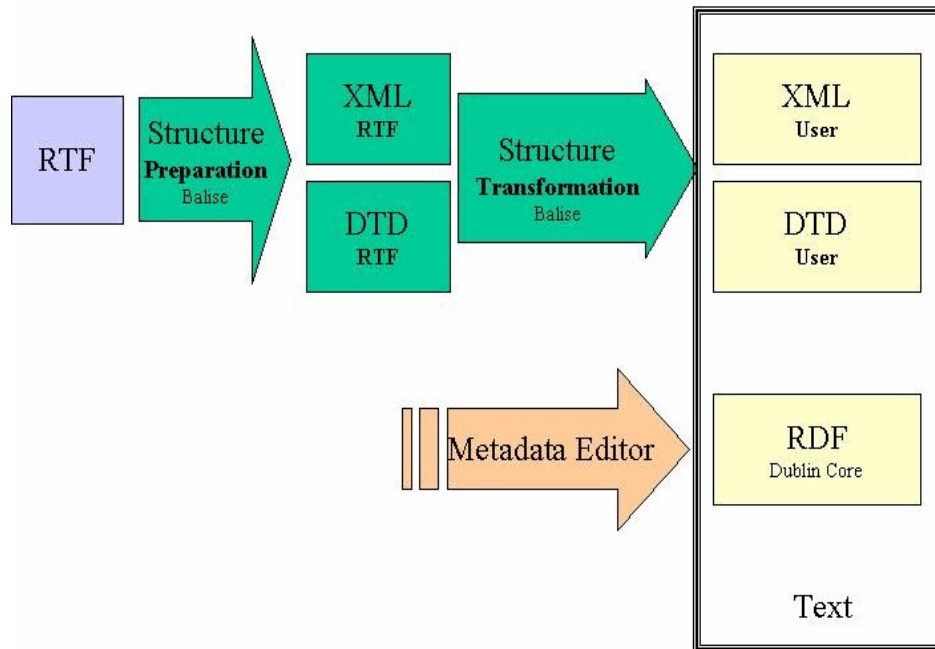


Figure 3. XML text component

Text must be written in Word 2000 in compliance with specific restrictions – i.e., pre-defined paragraph and character formats must be used - and text must be saved as RTF (Rich Text Format).

RTF documents are converted with a Balise transformation script using a typographically oriented RTF-DTD to XML instances conformant with the desired main structure of a given user DTD.

The resulting XML text documents do not contain any executable links. However, all places that may be related to links (potential anchors) are marked in the XML documents including necessary information for the link model.

In addition, for semantic description of the XML document RDF metadata (Resource Description Framework, Dublin Core) must be generated in XML syntax (special metadata editor).

Graphics component

VML (Vector Markup Language) and SVG (Scalable Vector Graphics) are both XML applications. The graphics described with VML or SVG lose their static property (pixels) and are used for displaying information intelligently as is also possible with CGM (Computer Graphics Metafile) .

A prototype system was developed in summer 1999 including CGM and VML to provide functions ranging from the generation of the participating components to processability and display options. The goal was to find a suitable format for integration in the XML system.

Graphical subcomponents can be selectively identified within an overall graphic and linked as units of information to one another and to text components. The text components can be XML elements from a text structured with a DTD or caption text taken from a graphic.

Based on a concrete example of an XML document, the prototype demonstrated the integrated use of VML and CGM:

- Scaling and navigation in VML and CGM components

- Separate manipulation of graphics and graphics text

- Linking of text components and graphics components

Microsoft Internet Explorer 5 was used for the display. All conversions were performed using the Balise transformation tool.

The ISO-Draw tool was used for CGM graphics; PowerPoint 2000 and VISIO 2000 were used for VML graphics.

The results of using VML and CGM were as follows:

- CGM can be shown using plug-ins (IE5 plus Netscape) with predefined scaling and viewing options.

- VML can be displayed directly in the Web browser without a plug-in.

- Thanks to a mathematical model, VML and CGM graphics are scalable.

- VML can be shown quickly in the browser.

- Usable VML can be generated with the aid of standard tools (Office 2000, VISIO 2000) and proprietary manipulations.

- Additional graphic formats, including DRW, WMF, EMZ and CGM can be converted to VML using suitable filters.

- SVG can presumably be used as an alternative without a great deal of additional expenditure.

- SVG and VML as XML applications can be integrated in an XML system more homogeneously than CGM.

The findings of the prototype and the premises had the effect of focusing document contents on the VML graphics format and on Microsoft Internet Explorer 5.

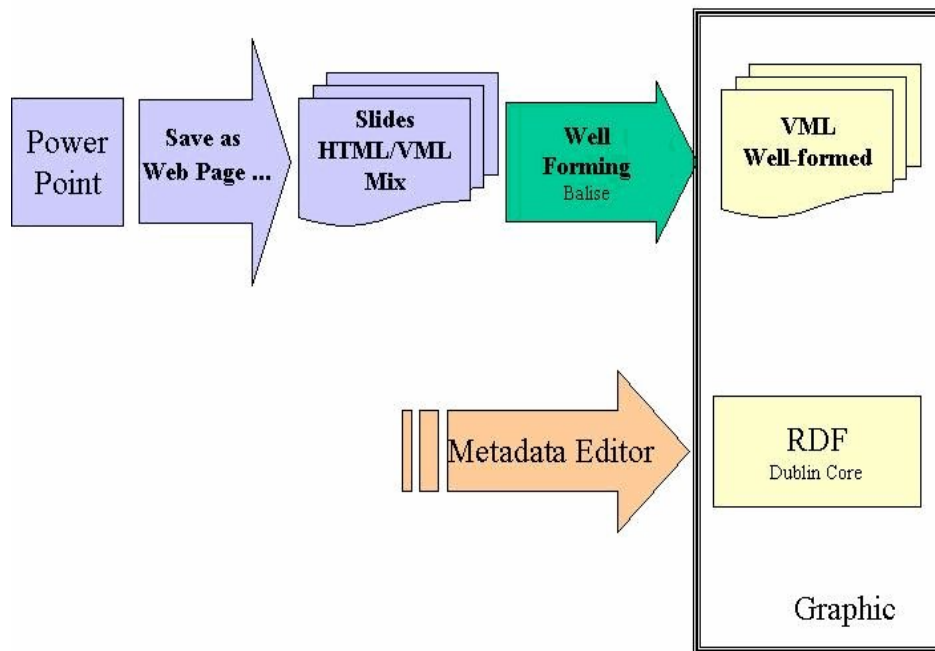


Figure 4. XML graphics component

Graphics referenced in a text document are compiled as a set of slides in PowerPoint. Each slide can be referenced as a graphic. The slides may contain different kinds of information: vector, pixel and text.

When saving slides in PowerPoint 2000 (Save as Web Page...) an HTML file is generated for each slide. In addition to the VML data the HTML files also contain text and image data. A set of slides generated in this manner is suitable for simple slide presentations using Microsoft Internet Explorer 5 but not for further editing and integration into the XML system.

Using a special balise transformation the HTML files were transformed into well-formed XML.

Graphics are to be regarded on the one hand as supplementary modules for structured XML text documents but, on the other hand, they can be used as individual graphic files. Consequently, as with text, the use of metadata (RDF) is imperative.

Link component

The use of an effective link model appears to be particularly important when integrating text and graphics.

The quality of linked information objects depends for the most part on how valid the links are – no sources without associated targets.

A common situation in the production of structured documents is the separation of content and typography. As a result, the same content may be represented in different ways. Breaking down documents into its constituent parts of contents, typography and link information grants authors and designers even greater flexibility.

Different types of links may occur in the link model used. Along with the simple unidirectional links, there are also bidirectional and multidirectional links. A link target can be a whole document or a particular location in a document. Specific descriptions for sources and targets of links in documents – so-called **potential anchors** – are inserted in the document contents during the production of text and graphics. In Word 2000 and PowerPoint 2000 the options for labeling links are used in compliance with relevant conventions to label potential anchors. A special element from the user DTD is generated for each potential anchor in the text instance, while a special attribute is generated in the graphic modules (VML) written in XML. Whether these labeled positions are finally used for links is determined in a subsequent processing step.

The fact that the author of a text and the graphics designer may well be different individuals should be regarded. Moreover, the times at which the individual components are completed may also differ. The same applies in particular for pure text information systems (joint editing).

It is recommended to separate link information from text and graphics. Links can therefore only be finally implemented once the text, graphics and link definition are available.

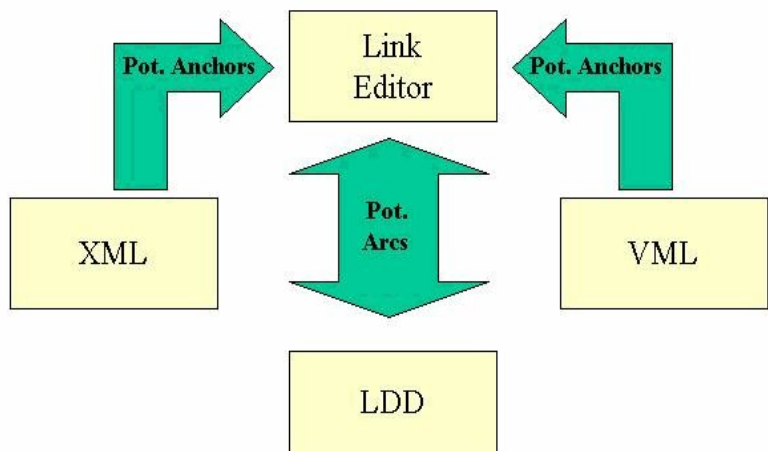


Figure 5. Editor - Link Definition Document

The link behaviour used is written using XML notation – the so-called LDD (Link Definition Document) – and is separately administered. The contents of an LDD are oriented according to a special DTD containing elements that define potential anchors. Secondly, the DTD contains elements for **potential arcs** representing the possible link connections. The entire link information in the XML system is written "out of line", while incorporating the principles of XLink/XPointer.

The actual document content (XML/VML) contains only locations for potential anchors. These locations are automatically extracted to the latest LDD during processing.

This link model allows individual, central administration of links (Link Editor). Before generating the Web data, all potential anchors can be extracted to the LDD from the documents involved in the generation process. Potential arcs can now be edited more easily with regard to all potential anchors.

It is, of course, also possible to first construct a system of potential arcs regardless of the availability of the actual associated text and graphics. This method of link designing can also be a basis for efficient processing and can be applied accordingly when assigning potential anchors.

Documents with identical content structures can share a link model (for example, subdocuments or language variants).

Visualization on the Web

Finally, when all necessary information is represented in XML, it can be basically used for generating multiple results. The main goal of the system however is to produce effective Web data enabling the use of new graphic and linking mechanisms.

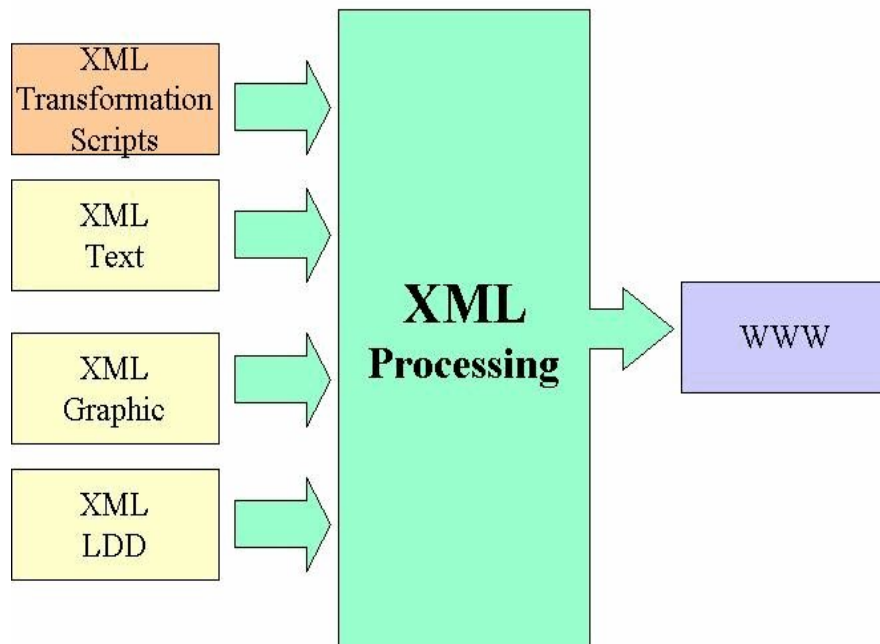


Figure 6. Generation process

An information and navigation system based on HTML is generated using a special balise transformation.

Results of the Web data generation:

- Corporate design of Web pages
- Detailed linking between text and graphics
- Valid link systems

Definable scaling options for graphics

Navigation in graphics

Provision of typography in the form of referenced CSS

Generated pages: index, glossary, table of contents, table of figures, etc.



Figure 7. Navigation

The Web data is based at present on HTML. Should the browsers and the XML standards prove to be sufficiently powerful, direct XML data solutions can be selected. Note, however, that the major benefit of XML is processing data. The support of different browsers makes generation variants necessary, which adversely affects the functionality.

Conclusion

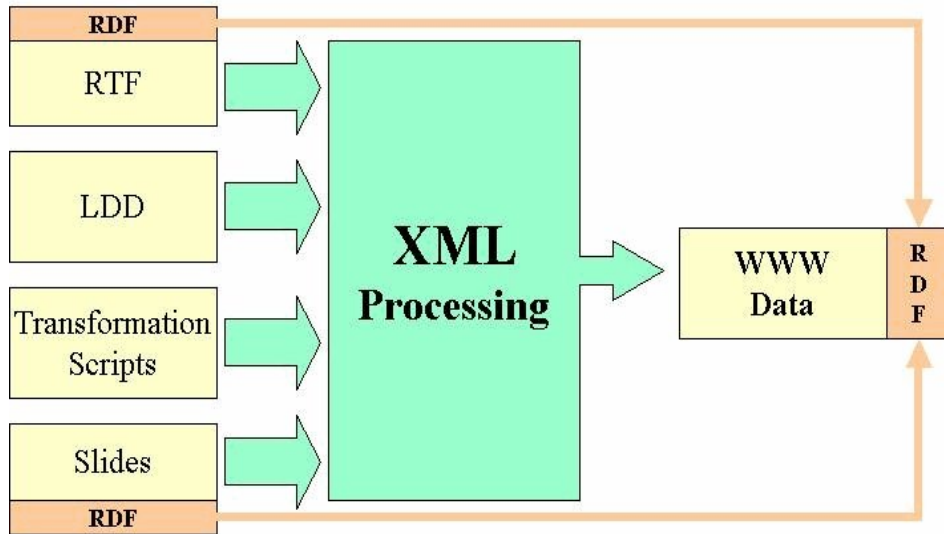


Figure 8. System

The decision to use XML is to be regarded as an evolutionary approach. The advantages of using XML-based graphics provide new aspects on the designing of information systems. The solution presented here must be adapted to the existing system environment. In this regard, questions of access security and the tasks of the document management system must be addressed. A wide application of the RDF model can be of great use for the future.

Bibliography

- [MKHParis98] C. Märtin, F. Hack, J. Krüger. Regulations Worldwide Online at the Siemens Public Communication Networks Group: SGML Editorial System for Providing Company-internal Regulations in the Intranet. GCA SGML/XML Europe '98 Paris.

Authors

Christian Märtin

Industrial Engineer Manager
Siemens AG, Information and Communication Networks, Central Organization
Postal Address:
Hofmannstr. 51
D-81359 Munich
Germany
Telephon: +49 89-722-48197
Fax: +49 89-722-34851
E-mail: christian.maertin@icn.siemens.de

Christian Märtin - Christian Märtin studied mechanical engineering and industrial engineering. Until 1972 he was System Engineer and Project Manager in development projects at "Motoren- und Triebwerks-Union (MTU)", Munich. Since 1972 Mr. Märtin has been in charge of projects in the information systems area at "Siemens AG", Munich. In collaboration with the Siemens Local Companies he set up a number of documentation centers in Latin America and since 1985 he has been doing foundation work in the area of structured information processing. Mr. Märtin is working on "Strategic Internet Concepts" in the Information and Communication Networks Group.

Jürgen Krüger

Business Consultant
Siemens Business Services, Structured Document Processing
Postal Address:
Carl-Wery-Str. 22
81739 Munich
Germany
Telephon: +49 89-636-54325
Fax: +49 89-722-34851
E-mail: juergen.krueger@mch.sbs.de
Web: www.sbs.de

Jürgen Krüger - Jürgen Krüger is a graduate in mathematics. From 1981 to 1992 he was Project Manager and Main Designer with responsibility for developing WYSIWYG editors at Berthold AG, Berlin. In 1992 Mr. Krüger moved to Siemens AG as a Technical Assessor. Today Mr. Krüger is a business consultant responsible for Document Management Systems and archives in the area of Structured Document Processing. Mr. Krüger has spoken at the following conferences: Fachtagung Technische Dokumentation '98 Munich, SGML/XML Europe '98 Paris.