Information Reuse in Multiple Forms

Covered topis: XML, CSS, XSLT, XSL-FO, CSS, ...

Executive Summary

This report explains the need for a computer to understand the documents that users create. By giving it this ability, a large problem known as information reuse will be solved. This problem was created because of the various new devices that were released. All of these devices need to view the same document; however, they all have different screen sizes and resolutions. Making the document presentable is achieved by a toolkit known as the Extensible Markup Language (XML).

There are many new technologies relating to XML. Throughout this document, the technologies will be compared, contrasted and analysed. The most important topics covered include the future of old documents, style sheets and document transformations. By confronting the information re-use problem, one can see the many new technologies that were spawned.

The new technologies discussed are not for everyday users. However, large companies will benefit from giving their customers a variety of ways to view the information they publish. A new and large market has emerged, and technology companies should strive to be on top of it.

1.0 Introduction

Today, desktop computers and paper documents are not the only way to view information. There are also Personal Digital Assistants (PDA), cell phones, electronic books (E-Books) and many new devices coming out every year. A document that was created on a desktop computer cannot be ported directly to all of the different platforms mentioned above.

This problem is generated in word processors, web page editors and many other authoring tools. They all intertwine presentation details along with the actual document data. For example, in a given document, the user might want to have the title appear in a large font, bold any text that has an emphasis, and write all quotes in italics. If this document is ported to a cell phone, the font will be too big, the colours in the document may not be supported and some sections of the text may not be needed. The user, then, has to create a second copy of the document with all these changes made.

The problem now becomes clear; several different copies of the same information will be present. Not only is space wasted, but if a spelling error is found, the user has to go through many different copies of the document and fix the error in each place. If the user supports the document on 10 different platforms, a lot of work will be generated. This document addresses this problem and introduces many new technologies created to accomplish this task.

2.0 Analysis

2.1The extensible markup language

Extensible markup language (XML) is the leading contender for the solution of information reuse. XML technology is simply metadata which is injected directly inside the original document text. XML identifies document elements such as titles, sections, chapters, important words and quotes. It is not responsible for doing anything more than this identification. XML was not meant to be used alone; instead, XML comes with a suite of other technologies that, when used together, can be very effective. XML is often referred to as this entire suite of technologies. This document will use the term XML more strictly, referring only to the embedded metadata in a document. It will also talk about an 'XML document' which is just a text document with XML metadata inserted inside.

An XML document was meant to be used with a style sheet. A style sheet simply describes how the sections of a document should be presented. In a broad sense, XML enables 1 sole copy of the actual document (XML document format) and several presentations style sheets. Once our original document text is marked up with XML metadata, users simply apply one of these style sheets, depending on the device that the document is being viewed on. To solidify an understanding of XML, one should see how XML looks. The figure below uses XML metadata to describe the different items of a grocery list.

```
<groceryList>
    <item>Apples</item>
   <item>Bread</item>
   <item>milk</item>
</groceryList>
```

Figure 1.0 - XML metadata

2.1.1 The World Wide Web Consortium

XML is not tied to any one company; it was created by the World Wide Web Consortium (W3C). The W3C develops specifications, guidelines, and tools which most major companies follow. They have created most of the internet standards that exist today, including but not limited to Hypertext transfer protocol (HTTP), Hypertext markup language (HTML), the document object model (DOM) and Cascading style sheets (CSS).

The W3C is composed of working groups, each of which take on a particular task and work towards a specific goal. Each working group is made up of key members from large organizations as well as well known individuals. The working group in charge of XML did not start developing its guidelines from scratch. Instead, XML is just a scaled down version of an existing very large and very complicated standard known as the Standard Generalized Markup Language (SGML).

2.1.2 The XML opportunity

Microsoft, Corel, Adobe, IBM and most other big technology companies are beginning to base a large portion of their technologies on XML.

The major reasons for this include

- **Reuse:** Separating the content of a document from the presentation details gives the ability to use the document on many different mediums.
- **Interchangeability:** Companies can use XML documents to communicate and share information with each other. Many XML-aware tools have been created to take advantage of this interchangeability.
- **Open standard:** By using XML, companies can create documents on any operating system with any tools. If a company ever decides to change the applications they use or the operating system they use, all of their documents can painlessly be transferred.
- **Simple editing:** XML can be edited with very simple tools such as notepad. XML is simply American Standard Code for Information Interchange (ASCII) text. ASCII text is the most popular standard way to describe computer text.

There are several other options a company or a user can use other than XML. SGML, one of the most obvious, hasn't become extremely popular simply because of its complexity. Most tools created for SGML now work just as efficiently with XML documents.

Another alternative to XML documents is a proprietary binary file format such as a

WordPerfect document. By using proprietary binary files, the four points mentioned above will be lost; however, ease of use will be gained. Binary file formats usually store presentation along with the content. They are not the best solution; however, they do have their place and reason for existing. Since the start of desktop computers, people have stored documents in the most efficient way possible. There was no need to store more information than what was needed. Programmers strived to use as few of the computer's resources as they could. Computers did not need to understand what a document meant, just as long as the users of the computer could. Today, much more powerful computers come with a variety of new devices. Computers now have the need to understand what a document means. By having this ability, computers can be intelligent enough to know how to format a document depending on the device the document is being viewed on. Computers have also gained the ability to do complicated things with these documents. Listed below is a table describing the differences between XML, SGML and binary file formats. From this the user can clearly see why XML has been chosen as the industry standard of today, and why the binary file format has been the industry standard up to this point.

Capability	**XML**	**SGML**	**Binary file**	**Importance today**
Small file size	-	-	+	-
Ease of use	+	-	++	+
Availability of several tools	+	+	-	+
Reuse	+	+	-	+
Interchangeability	+	+	-	+
Open standard	+	+	-	+

Editable in an ASCII text editor	+	+	-	-

Figure 2.0 - XML, SGML and Binary file format comparison

2.2 Style sheet descriptions

So far, this report has established that presentation and actual content should remain separate. Since XML only identifies the different parts of documents, one must also use style sheets to convert the XML documents into a pleasing readable format. This section will describe and contrast the technologies for adding style to documents.

The most widely used style sheet format today is called Cascading style sheets (CSS). CSS was created to be used along with XML. This format has been widely accepted and used for several years. It is very simple to use and describes a document's font, colour and layout. CSS was also created by the W3C.

Recently, the W3C noticed the need for something more than CSS. The W3C created the extensible style sheet language formatting objects (XSL-FO) to satisfy these needs. XSL-FO extends presentation one step further. A document can now be converted to an audio file and read back, a video file that can be animated and watched or even converted to braille. With the help of style-sheets, creativity is limited only by imagination.

2.2.1 XSL-FO is not always best

Both of these standards are ‘recommendations’ by the W3C. However, even W3C recommends the use of CSS before XSL-FO. The reason for this is that CSS is easier to use, learn and maintain.

As the below figure describes, XSL-FO is actually another form of XML. Therefore any tool developed for XML parsing and processing can be used with XSL-FO. Overall most of the XML community foresees XSL-FO as being the dominant style sheet language in the future.

Capability	**CSS**	**XSL-FO**
Data separate from style	+	+
On screen and printed positioning	+	+
Colour support	+	+
Font size	+	+
braille	-	+
audio	-	+
video	-	+
Easy to use	+	-
Ability to add/remove actual content of the document	-	+
Is itself an XML format	-	+
In popular use today	+	-

Figure 3.0 - CSS VS. XSL-FO

2.3 Document transformations

Suppose a document is created using XML for all the benefits mentioned up to this point. One may feel the need to extract all the quotes that occur in this document, or perhaps extract all the titles of each section to create a table of contents. The only way up until now would be to copy and paste each section into a new document.

The original problem is beginning to recreate itself. Two documents which contain a big chunk of common data would be present. To make changes to a titles in a document, a change it in two different places would be required. The W3C has resolved this problem by creating the extensible style sheet transformation language (XSLT).

2.3.1 XSLT pros and cons

XSLT transforms your document from one form to another. This transformation resides in an independent XML file from the original content XML file. Both the transformation XML file and the content XML file are used together in a translation program such as *XT by James Clark*. The output is a third XML file with the applied transformations.

The three typical uses of XSLT are:

1. To generate a new document which contains only a subset of the original document.
2. To generate a new document which contains the original document along with extra information
3. To transform a document from one form to another. For example, different

companies may use different XML file structures to describe the same data. These companies can interact by having 2 style sheets that will transform their document from one format to the other.

XSLT is the ideal solution to the concept of derived documents. It is also the leading technology that drives company interaction with XML files. Learning XSLT is a big task and is equivalent to learning the syntax of any other programming language.

2.4 XML output, much more than text

The W3C, several companies and several individuals have created standard XML format. An XML application is an XML document which is constructed in a predefined structure. All XML documents are written in ASCII text, however, the programs that interpret the XML document may have an output other than text.

Today, XML formats to describe text, music, vector graphics, math and much more exist. Not all XML document interpreters need to be used to provide a visual presentation to the user. For example, a software developer may create 3 separate software programs for the music XML applications. The first might display the described music as sheet music, the second actually play the music, and the third give the ability to transpose the document into any key. All of this can be accomplished by only having one XML application created to describe music. This is a great improvement from past methods which would require 3 separate binary file formats.

At this point one should be able to see how powerful it can be for a computer to understand what a stored document means. Many standard XML applications (structured XML documents) have been created. Programmers around the world can create programs to work with these XML applications and each program can potentially do something completely different. For example, once all of Mozart's work has been described in XML documents, they can be used by all of the music programs described above.

2.4.1 XML applications

Several XML applications are widely used today. An XML application follows a predefined layout. The definition of this layout is created with something called an XML schema. The number of XML applications (and therefore schemas) are continuously growing as XML proves itself to be the ideal solution to business problems around the world.

Some of the more popular XML applications include:

- **The Math Markup Language (MathML):** Working with mathematical equations used to be a big pain. MathML describes any mathematical formula in an easy to understand syntax. By using MathML, one program might draw the graph of an equation wile another program simply displays the math equation. A third program can display and allow the manipulation of the equation.
- **Scalable Vector Graphics (SVG):** The ability to create simple graphics with a small file size. This works by instead of storing the picture, storing a description

of the picture. One program might display the picture while another allows the picture to be edited.

- **Synchronized Multimedia Integration Language (SMIL):** Allows the creation of animations and presentations. SVG is often used in conjunction with SMIL.
- **Extensible Hypertext Markup Language (XHTML):** Reformulation of HTML 4.0 to follow the strict guidelines of being an XML document

By solving our original simple problem, a whole set of tools for creating every type of multimedia imaginable has been introduced. Most XML applications can be used collaboratively since they're all based on the XML metadata.

3.0 Conclusion

XML describes the different parts of a document with metadata. Instead of saying that a title should appear in a large font, the title can be marked with metadata as actually being a title. This metadata enables the use of style sheets, transformations and parsing of the document. These technologies, created by the W3C, give us the ability to solve a long standing problem: having multiple documents for multiple platforms containing the same data.

By understanding and marking the various parts of the document, it can be converted braille, audio and video. People with disabilities will now have complete and unrestricted access to all the information that available. It is extremely important to allow our computer to understand our documents.

4.0 Recommendations

For simple everyday tasks, simple text editors such as Word perfect, Word or HTML editors are recommended. The everyday user need not be concerned with anything more complicated. However, publishers, and companies alike definitely will find those limitations to be very transparent. For this class of people, It is recommended that they describe all their documentation and information in the XML format. The XML format is widely used, it gives us many abilities over binary file formats and it is much easier to use than SGML. Regarding document presentation, CSS is recommended when it can be used. The use of XSL-FO is better restricted to an “on need” basis. Document transforms should never be done by hand; instead the use of XSLT is beneficial.