



Genealogical Computing Group

Special Interest Group of the
New Zealand Society of Genealogists
Southland Branch



XML – February 2009

What is HML? Markup language is a set of codes that give instructions regarding the structure of a text or how it is to be displayed. A well-known example of a markup language is HyperText Markup Language (HTML), one of the most used in the World Wide Web. For example, if an HTML page contains "test" a typical Web browser will display the word "test" in bold face; the "" and "" won't be displayed because, as markup, they are instructions to the browser, not part of the content.

What is XML? XML (Extensible Markup Language) is a general-purpose specification for creating custom markup languages. It is classified as an extensible language, because it allows the user to define the mark-up elements. XML's purpose is to aid information systems in sharing structured data by encode documents.

Why use XML? Governments and large organizations around the world are becoming very conscious that their electronic documents are locked away in file formats that can no longer easily be read, and it is an expensive exercise to get them converted to a modern format. This problem is ongoing while proprietary file formats are used. Open Office uses a XML format for all its data storage, and to address Government concerns and possible loss of clients Microsoft has also adopted XML with its latest version of Office. Open Office .odt and MSWord .docx files are both compressed XML files. If you add .zip to the file name (no need to remove the existing extension – it just becomes part of the file name) the files can be opened as a normal zip file and the contents viewed as plain text. Microsoft and Open Office use different XML definitions, and as Open Office is simpler I will use it to demonstrate.

The first thing is that there are a lot of small files and subdirectories in the document. The only one we will look at is "content.xml" ("document.xml" in MSWord). The rest of the files store definitions of styles, fonts etc. which are part of the XML definition.

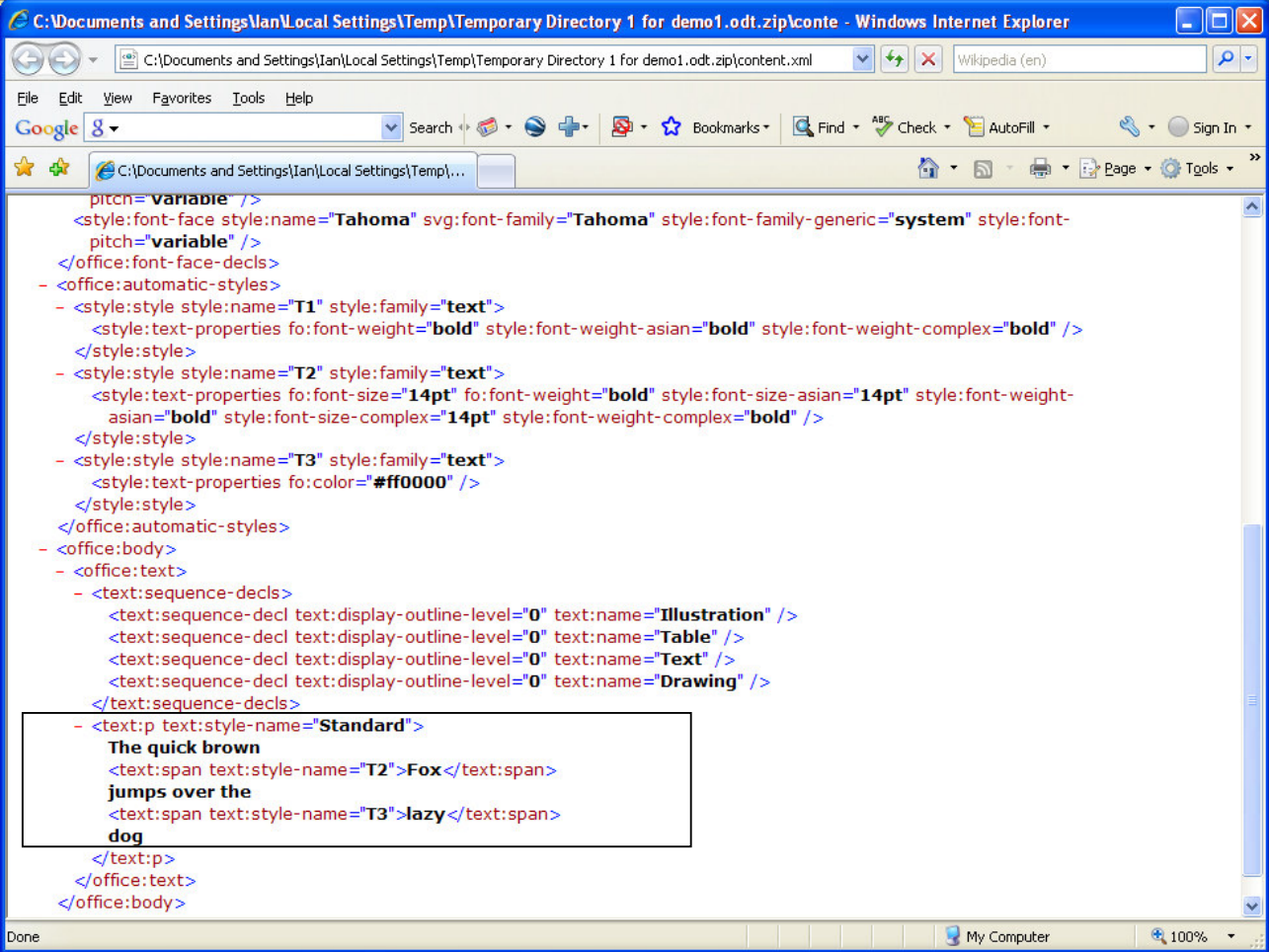
The screenshot shows a Windows Explorer window titled "C:\My Documents\lan\Genealogy\demo1.odt.zip". The address bar shows the path "C:\My Documents\lan\Genealogy\demo1.odt.zip". The main pane displays a list of files and folders:

Name	Type	Package...	Has ...	Size	R...	Date
Configurations2	File Folder	0 KB		0 KB	0%	
META-INF	File Folder	0 KB		0 KB	0%	
Thumbnails	File Folder	0 KB		0 KB	0%	
content.xml	XML Document	1 KB	No	4 KB	76%	23/02/2009 10:12 p.m.
meta.xml	XML Document	1 KB	No	1 KB	0%	23/02/2009 10:12 p.m.
mimetype	File	1 KB	No	1 KB	0%	23/02/2009 10:12 p.m.
settings.xml	XML Document	2 KB	No	8 KB	85%	23/02/2009 10:12 p.m.
styles.xml	XML Document	2 KB	No	11 KB	83%	23/02/2009 10:12 p.m.

The left sidebar shows "Folder Tasks" and "Other Places". The "Details" pane shows information for "demo1.odt.zip":
Compressed (zipped) Folder
Date Modified: Yesterday, 24 February 2009, 11:12 a.m.
Size: 7.66 KB

The status bar at the bottom indicates "8 objects".

If you click on content.xml the xml will display in the web browser. This example is the xml to display "The quick brown **Fox** jumps over the *lazy* dog". T2 is defined as 14pt and bold, while T3 is defined as red (#ff0000). I have put a box around the actual text to make it easier to follow.



- This instructs the computer to display the following:-
- > Standard format The quick brown
 - > 14pt bold (T2) Fox
 - > Standard format jumps over the
 - > Red (T3) lazy
 - > Standard format dog

You can see that if you use a large number of different fonts and sizes your document rapidly increases in size.

This is a part of a Word .doc file of the same text viewed in notepad (no indication of formatting). The gaps and strange characters are all hexadecimal numbers which tell Word, but not the viewer, how to display the text. This file is twice the size of the XML files.

