XML vs. HTML: A Publishing Comparison

for the
United States Bureau of the Census's
Statistical Compendia Branch

Prepared by

http://www.chm.net/

4710 Auth Place
Suite 590
Camp Springs, MD 20746
Voice: 301-899-2601
Fax: 301-899-1555

July 19, 2002

1.

Executive Summary

On August 7, 1998, then-President Clinton signed into law the Rehabilitation Act Amendments of 1998. This Act covers access to federally funded technology-based programs and services, and requires access to electronic and information technology procured by the Federal government be provided to persons with disabilities, to the extent that this access does not pose an "undue burden." On December 21, 2000, the Architectural and Transportation Barriers Compliance Board (Access Board) issued final accessibility standards for electronic and information technology covered by Section 508 of the Rehabilitation Act Amendments of 1998 (Section 508). These standards speak to various means for disseminating information, including computers, software, and electronic office equipment. The law also provides a complaint process under which complaints concerning access to technology are to be investigated by the responsible Federal agency.

The United States Bureau of the Census' (Census Bureau) Statistical Compendia Branch regularly produces several popular statistical reports that are made available as Adobe Acrobat PDF files on the United States Bureau of the Census' Web site at

http://www.census.gov/statab/www/. The product line includes the *Statistical Abstract of the United States*, and the *State and Metropolitan Area Data Book 1997-98*.

The Census Bureau has contracted Computer & Hi-tech Management, Inc., (CHM) to conduct an assessment of these PDF files as regards Section 508 and their compliance with same, measured against the December 21, 2000, Final Rule published in the Federal Register, and paying particular attention to the Technical Standards published for Subpart B, § 1194.21 and § 1194.22, and Subpart C, § 1194.31.

This paper evaluates the alternatives of using HTML (HyperText Markup Language) or XML (eXtensible Markup Language). It concludes that the XML approach offers greater flexibility for the future.

Table of Contents

1. Document Scope

This evaluation compares the use of HTML (HyperText Markup Language) against the alternative of XML (eXtensible Markup Language). These are two competing technologies for storing text and displaying it in a variety of formats.

2. Introduction

The Rehabilitation Act Amendments of 1998 requires access to electronic and information technology procured by the Federal government be provided to persons with disabilities, to the extent that this access does not pose an "undue burden." On December 21, 2000, the Architectural and Transportation Barriers Compliance Board (Accessibility Board) issued final accessibility standards for electronic and information technology covered by Section 508 of the Rehabilitation Act Amendments of 1998 (Section 508). This covers computers, software, and electronic office equipment. Complaints concerning access to technology are investigated by the responsible Federal agency.

The Statistical Compendia Branch regularly produces several popular statistical reports as Adobe Acrobat PDF files on the United States Bureau of the Census' Web site at

http://www.census.gov/statab/www/. The product line includes the *Statistical Abstract of the United States*, the *State and County Area Data Book*, and the *County and City Data Book*.

In keeping with Section 508, all Census Bureau Web pages, inclusive of PDF files, published after July 21, 2000, must be made accessible to persons with disabilities (*i.e.*, visually-impaired and blind persons). Additionally, the United States Department of Commerce established a goal of making *all* of its Web data products fully accessible to persons with disabilities. The Section 508 accessibility requirements for these documents are two-fold:

1. The documents must be accessible to persons with visual impairments, *i.e.*, those who are "hard of seeing," who are color blind, or who have limited contrast-perception, etc.; and

2. The documents must be accessible to persons who are blind or who have sufficient visual impairment such that they use assistive devices, *e.g.*, screen magnification software, text-to-voice screen readers, and refreshable Braille displays, when utilizing the Web.

Additionally, Section 508 requires Federal agencies ensure accessibility internally to Federal employee as well as to the general public. Any technology procured for the purposes of making the subject Adobe Acrobat PDF files compliant with Section 508 must, in turn, also meet the accessibility standards laid out therein. In this instance, Adobe Acrobat 5.0, if procured by the Census Bureau to be used by Census Bureau employee, must by a Section 508 compliant software package.

Adobe, Inc. currently markets Acrobat Reader 5.0 and Acrobat 5.0, which provide a number of new capabilities that improve the accessibility of both the software and the information communicated in PDF files.

On October 11, 2001, Computer & Hi-tech Management, Inc. (CHM) assessed the Statistical Compendia Branch's Web publications and the Branch's compliance with Section 508. CHM made a number of recommendations, specifically:

1. The Census Bureau provide a link to Adobe's free conversion service for its PDF files;

2. The Census Bureau use Adobe Acrobat 5.0, along with the various accessibility-enhancing plug-ins offered by Adobe Systems, to increase the accessibility of its PDFs;

3. The Census Bureau provide custom created tables for individuals requesting them;

4. The Census Bureau provide tables using the new HyperText Markup Language (HTML) 4.01 accessibility enhancements; or

5. The Census Bureau use eXstensible Markup Language (XML) to generate tables.

The Statistical Compendia Branch charged CHM with exploring in greater detail the last two options, analyzing and comparing the costs and benefits for both choices.

3. Section 508 Accessibility Requirements: A Review

Section 508 addresses accessibility standards for electronic and information technology programs and services procured or provided by the Federal government, and seeks to guarantee that persons with disabilities have equal access to those programs and serves as do person without disabilities. Most observers correctly understand this to mean that Federal Web sites must be accessible to persons with disabilities.

Section 508, however, additionally states that third-party plug-ins, applets, or other applications utilized to view or interact with content deployed on Web sites must also be compliant with the requirements of Section 508, and requires that Web content authors provide a link to that plug-in, applet, or application. *See*, 36 CFR part 1194, Section 508 of the Rehabilitation Act Amendments of 1998, at § 1194.22 (m).

Finally, Section 508

> "requires that when Federal agencies develop, procure, maintain, or use electronic and information technology, they shall ensure that the electronic and information technology allows Federal employees with disabilities to have access to and use of information and data that is comparable to the access to and use of information and data by Federal employees who are not individuals with disabilities…." Fed. Reg. *Electronic and Information Accessibility Standards*. Vol. 65, No. 246, Rules and Regulations, pp. 80500 to 80528. at 80500.

Specifically, Section 508 appears to require:

1. All data (including PDFs) published by the Statistical Compendia Branch on the Census Bureau's Web site be accessible;

2. Any plug-in programs such as Adobe Acrobat Reader be accessible to persons with disabilities; and

3. Products such as Adobe Acrobat Writer used by the Census Bureau to publish data on the Web be accessible themselves.

Three Subparts of Section 508 are immediately relevant to the Census Bureau's publication of Adobe PDF documents online. They are:

1. Subpart B, § 1194.21, which governs software applications and operating systems. Regulations listed under § 1194.21 are relevant to the Bureau's assessment of Adobe Acrobat 5.0 and Adobe Acrobat Reader 5.0, and those program's compliance with accessibility

requirements;

2. Subpart B, § 1194.22, addressing Web-based Intranet and Internet information and applications. Regulations listed under § 1194.22 speak directly to the Census Bureau Web pages; and

3. Subpart C, § 1194.31, which define the functional performance criteria by which an information technology is to be measured.

4. Current Publishing System

The Statistical Compendia Branch currently uses a hybrid system that starts with a Lotus spreadsheet for each table that is to appear in the *Statistical Abstract*. The spreadsheet has a column with special row formatting information. Each table is saved as an ASCII file. Once several ASCII fields are created, they are collated and sent to XYVision where they are converted to PostScript files for hard copy publication and to PDFs for electronic publication on the Web and on a CD-ROM. A copy of the Lotus spreadsheet is also put on the CD-ROM (without the special formatting information).

This process is very labor intensive because the data for the tables come from a wide range of government agencies, international sources and domestic groups. Each year each table must be updated, edited and verified. A spreadsheet is critical because it allows the branch to verify that the column totals and subtotals are correct. Most tables have some information dropped each year to make room for the new data and to prevent the *Statistical Abstract* from becoming too large to publish in paperback.

Recently XyEnterprise announce that it would no longer support the version of XYVision used by ACSD. XyEnterprise and Interwoven have created a partnership that will replace XyVision with a new product to be called XPP. This will provide automated XML to PDF integration

5. Requirements

The Statistical Compendia Branch has a number of requirements for a publishing system:

1. It must be able to target paper, Web and CD-ROM publishing.
2. It must be able to easily import data from a variety of sources outside of the Census Bureau.
3. There must be a convenient way to double-check the tables to make sure that entries are accurate.
4. It should be flexible to adapt to future technologies.
5. It should not require extensive retraining of branch staff.
6. It should make access to the branch's Web pages accessible to the visually handicapped at least at the Section 508-level if not better.

In a previous report, the alternatives were reduced to using HTML and XML.

6. Comparing HTML And XML

Both XML and HTML are derived from the more complex Standard Generalized Markup Language (SGML). Programs based on SGML are very complex and expensive. As a result, SGML has not been widely adopted.

HTML is the most widely used markup language for Web-based documents. A document using HTML contains embedded tags that provide guidance to HTML viewers (usually called Web

browsers) as to how to display the document and connect it to other documents. HTML provides suggestions for the display of text, although in recent years tools have been developed to specify the appearance as rigidly when seen in a Web browser as on a printed page. HTML has three key attributes:

- **Linkability:** Data is (hyper)linked in HTML, letting one piece carry you to another.
- **Simplicity:** HTML is simple, making it easy to learn and to display.
- **Portability:** HTML is stripped down, making it portable over networks, operating systems and languages.

However, as the popularity of HTML has increased, the limitations of the language have become more apparent. HTML is not extensible. HTML authors cannot add new features to the language. The language definition prohibits this and commercially available Web browsers ignore tags that are not part of the HTML standard. To make matters worse, the main browsers (Microsoft's Internet Explorer and Netscape's Navigator) have unique tags that extend HTML and ignore some standard tag. The tags that control presentation are in the same file with tags that describe the document content. Although HTML 4.01 and Cascading Style Sheets (CSS) enable HTML authors to separate content from presentation, HTML 4.01 remains weak in its ability to describe the content of a document.

HTML 4.01 added a number of tags to tables to allow non-displaying comments to be added. Screen readers such as JAWS read the information in these tags to vocalize additional information for the visually impaired. These tags include a table summary, cell, row, row group, column and column group comments. A fully commented table would use all of these and someone using a screen reader could easily be overwhelmed by information.

Finally, while it's relatively easy to create a Web page, it's far more difficult to maintain a large Web site published using "flat" HTML files. Many large Web sites today use database publishing, dynamically generating pages using Java Server Pages, Active Server Pages, and other middleware products connected to a database in which content is stored.

HTML markup is fixed, as linkability, simplicity, and portability demand limits on markup. However, HTML is limited in:

- **Intelligibility:** How well data knows itself.
- **Adaptability:** How well data changes in response to environmental changes.
- **Maintainability:** How easily data is maintained.

XML overcomes limitations of HTML and other markup languages, while providing capabilities that are not a part of the earlier languages.

Like HTML, XML makes use of a tagset, but while HTML specifies what each tag and attribute means (and often how the text between them will look in a browser), XML uses the tags only to delimit pieces of data, and leaves the interpretation of the data completely to the application that reads it**.** XML is a *meta-language*, or a language used to define other languages. As such, XML can be used to define HTML. In an XML document, the tag names convey the meaning of the data they contain. Document structure is easily discernable and follows a pattern. In contrast, HTML tag names reveal little about the meaning of their content and the structure is not particularly useful for manipulating the document and exchanging it between applications.

This is especially important because, in order for a document to communicate successfully from author to reader, both parties must agree that words mean what all choose them to mean. *Semantics* can only be interpreted within the context of a community. Millions of HTML users worldwide, for example, agree that **<B>** means bold text, or that **<H1>** is a prominent top-level document heading. The same cannot be said, though, for the date *8-7-97*, which reflects local culture, or for **<font face=Arial>**, which is only usable by Microsoft Windows systems. The larger the community, the weaker the shared context; the smaller and more focused the community, the stronger the shared context becomes.

HTML is currently the only common tagset Web users can rely on. Furthermore, HTML cannot be extended unilaterally, since the shared definition is maintained by a central standardization organization that publishes new versions of the markup language (*i.e.*, HTML 2.0, HTML 3.2, and HTML 4.0). Since semantics depend on shared agreements between readers and writers about the state of the world, there is a place for community-specific definitions. XML makes it cost-effective to capture community ontologies as Document Type Definitions to decentralize the control of specialized markup languages. The emergence of richly annotated data structures catalyzes new applications for storing, sharing, and processing ideas.

XML is designed to be flexible. If accessibility requirements change in the future, XML will be able to meet them.

However, XML is considerably more difficult to use. XML is not well supported by browsers today so it cannot replace HTML on a Web page, but it can be processed on the server to create an HTML page. Netscape and Internet Explorer do not fully implement XML and the related eXtensible Stylesheet Language (XSL). Their implementations are inconsistent. They are changing with every new release of Netscape and IE make Web design using XML difficult.

The industry solution has been to use XML on the server side and to use additional software to transform it to the HTML that any Web browser can use. An additional advantage to this server side approach is that if a Web site is designed to be used by Portable Digital Assistants (PDAs) such as Palms and Compaq iPAQs, the computational burden is place on the vastly more powerful server.

The server side solution can also reduce the amount of information that is sent to the browser. XML is also verbose. Since XML is a text format using tags to delimit data, XML files are nearly always larger than comparable HTML formats. This can improve performance over slow connections such as found with wireless access.

6.1. Using HTML to Publish the Statistical Compendia Branch's Data Products

The Branch could replace the current PDF tables with HTML tables. This could be done manually or with a new automated system.

The advantage of this is that there is one presentation for all users. By using an open standard, the burden of correcting any problems in accessing data would not fall on the Census Bureau. The result of this effort is to update the Web page, but the HTML could also be stored on the CD-ROM.

The disadvantage is that manual table conversion would be subject to errors and is slow. To automate the process a new entire workflow would have to be developed. The original tables must be transformed into a format that can produce HTML 4.01 tables without preventing it from being used for the hard copy publication of the Statistical Abstract. The cost is likely to be substantial.

The non-display tags would have to be updated each year.

6.2. Using XML to Publish the Statistical Compendia Branch's Data Products

In this alternative, the Census Bureau would mark-up its tables with customized XML tags that accurately describe tabular and data relationships. This eXtensible mark-up would provide assistive technologies with information sufficient to allow a disabled person to know that he was, for instance, reviewing data in a table containing spanning column headers and complex parent-child row headings, such that he could correctly associate data and metadata irrespective of the table's visual presentation.

XML is the natural technology to address the specific parameters of the Statistical Compendia Branch's business model: a multitude of diverse assets undergo constant revision and transformation in a multi-user environment. Because of its reusability, extensibility, and modularity XML plays a critical role in the successful management of these assets. Moreover, content captured in XML can be reused in all facets of a business: It can be re-purposed for multiple delivery channels; It can be aggregated with other content based on presentation context; or it can be put to many other uses. Ultimately, by leveraging XML, the Branch can make the most of their content.

Additionally, XY Enterprises, the service provider that provides publication support to the Statistical Compendia Branch, has integrated XML into its product service offerings, and now supports both Web delivery and traditional print output. Further, XY Enterprises recently announced a partnership with content management software provider Innervision that further integrates XML. The incorporation of XML into Web content management could easily be integrated with the Statistical Compendia Branch traditional publication role.

However, as discussed earlier, XML is a significantly more difficult markup language to learn and master; as such, the Bureau would need to invest some significant "human capital" in acquiring a knowledge base sufficient to implement an XML-based content management and publication system.

XML could be integrated into the current process by a combination of tools based on the existing Lotus worksheets. One approach is to bracket each data column with a pair of XML tags. Most of the time, the tag columns can be hidden facilitating editing the worksheet (Illustration 1).

Illustration 1—Tags Hidden

13	reservoirs, draining of lakes, etc. Density figures are					
14	based on land area measurements as reported in earlier censuses]					
15						
16	----------	------	----------	------	----------	----------
17					RESIDENT POPULATION	
18		------	----------	------	----------	----------
19					Per	
20	CENSUS DATE				square	Inc
21			Number		mile of	prece
22					land area	----------
23						Numbe
24	----------	------	----------	------	----------	----------
25	CONTERMINOUS U.S. \1					
26	1790 (Aug. 2)		3,929,214		4.5	
27	1800 (Aug. 4)		5,308,483		6.1	1,379
28	1810 (Aug. 6)		7,239,881		4.3	1,931
29	1820 (Aug. 7)		9,638,453		5.5	2,398
30	1830 (June 1)		12,866,020		7.4	3,227

When necessary, the columns are unhidden to reveal something similar to Illustration 2.

Illustration 2—XML Tags Visible

26	<YEAR>	1790 (Aug. 2)	</YEAR>		<ResPop>	3,929,214	</ResPop>
27	<YEAR>	1800 (Aug. 4)	</YEAR>		<ResPop>	5,308,483	</ResPop>
28	<YEAR>	1810 (Aug. 6)	</YEAR>		<ResPop>	7,239,881	</ResPop>
29	<YEAR>	1820 (Aug. 7)	</YEAR>		<ResPop>	9,638,453	</ResPop>
30	<YEAR>	1830 (June 1)	</YEAR>		<ResPop>	12,866,020	</ResPop>
31	<YEAR>	1840 (June 1)	</YEAR>		<ResPop>	17,069,453	</ResPop>
32	<YEAR>	1850 (June 1)	</YEAR>		<ResPop>	23,191,876	</ResPop>
33	<YEAR>	1860 (June 1)	</YEAR>		<ResPop>	31,443,321	</ResPop>
34	<YEAR>	1870 (June 1)	</YEAR>	<SuperScript>2</SuperScript>	<ResPop>	39,818,449	</ResPop>

7. Conclusion

The cost of converting to a modern system is not as great as it might seem because the cost of not doing anything is also significant. First, XYVision is shortly dropping support for the version being used by the Statistical Compendia Branch. Without vendor support, resolving new problems will become very time consuming and frustrating. To continue to be able to purchase vendor support will require an upgrade to XML Publisher. Second, the new version offers a host of new features that address the issues of unified publishing to various media and accessibility that concern the Statistical Compendia Branch. Third, by adopting an XML-based solution such as XML Publisher, the Branch places itself in the mainstream of document technology where it can take advantage of new developments most easily and enhance its mission to make information available to the public.