International Journal of Information Technology and Knowledge Management
January-June 2012, Volume 5, No. 1, pp. 123-128

AN ANALYSIS IN COMPARISON RELATED TO THE PROBLEM OF DEVELOPING WEB-BASED INFORMATION SYSTEMS

Sudesh Kumar[1], and (Mrs.) Ela Kumar[2]

ABSTRACT: This paper aims to examine the contextual issues relating to the problem of developing web-based information systems for emergent organizations. World Wide Web, as a public source of information, contains enormous amount of data in different formats. This paper critically examines four web-based information storage formats and the searching mechanisms being used in them. The scope of this paper is limited to HTML, XML, PDF and Image format. It describes briefly the leverages and perils of using each of these formats and critically compares their support systems. The paper also offers advice to Librarians to enable them to manage web-based information in a better way for their users by using suitable format(s).

Keywords: HTML, XML, PDF, Image formats, data formats

1. INTRODUCTION

The overwhelming mass of available information and its ever increasing number of characteristics have made it really difficult for librarians to preserve, search and retrieve information pin-pointedly, exhaustively and expeditiously. One answer to the problem could be digital preservation. Digital preservation concerns itself with ensuring that the records which are created electronically will remain available, usable, and authentic in ten to one hundred years time, when the applications and systems which were used to create and interpret the record will, more likely than not, no longer be available. Digital preservation consists of preserving more than just the record's bit stream. During preservation, questions of record context, content, structure, appearance, and behaviour must also be taken into account. Appearance and behaviour are aspects that are peculiar to digital records. These may, therefore, require the most attention to authentically preserve the record over the long term. There is a wide range of digital formats available, and to make matters more complicated, different digital objects have different preservation requirements. These days, "HTML" "XML" "PDF" "Image" formats are most popular standardization efforts in web documentation/information representation, and are rapidly becoming a standard for data representations, searching and exchange over Internet. This paper critically examines on the use of web-based information/documentation storage formats such as HTML, XML, PDF, Image format and their searching parameters.

The rest of the paper is organized as follows. Sections II, III, IV and V will describe about various technologies (i.e. HTML, XML, PDF and Image Files) which can be used to develop WBIS (Web-Based Information System).In section VI, we will present how searching is done in these technologies. Section VII presents a comparative analysis of these technologies. Finally, Section VIII will conclude the discussion with possible future scopes.

2. HTML

If information has to be stored on a central computer, it must be created first. While being created, information can be stored in the different forms and stored as files on the computer. These files are created using special software programming environment. Some forms/formats of files are HTML, XML, PDF and Image format etc. Files that travel across the largest network in the world, the Internet, and carry information from a "server" to "client" that requested them are called "webpage". The language used to develop webpage is called Hyper Text Mark-up Language (HTML).

3. XML

Extensible Mark-up Language (XML) is a simple, very flexible text format derived from SGML (ISO 8879). In 1998, it was published as an open standard by the World Wide Web Consortium (W3C). Originally designed to meet the challenges of large-scale electronic publishing, XML is playing an increasingly important role in the exchange of a wide variety of data on the web and elsewhere. XML is a meta-language that is a language that describes a language, which can be used to define an infinite number of customized mark-up languages. XML specification defines the syntactic rules governing its usage, the element, or tags, used within XML are created by its users. This ability to

[1] Associate Professor Department of Information Technology, BRCM CET, Bahal (Bhiwani),
[2] School of Information and Communication Technology Gautam Buddha University Greater Noida (U.P.)
E-mail: sudeshjakhar@gmail.com

create and define elements is the extensible aspect of XML. Like SGML, XML elements and their relationship are defined in a Document Type Definition (DTD). A program called a Parser can be used to check that the XML document is valid according to the rules defined in the DTD. XML is a structural mark-up language. XML also has the means to create hyperlinks to various kinds. It is relatively simple to use. XML is a framework for defining document mark-up languages. In simple terms, a document mark-up language is a set of element (frequently called tag) that has one or more of the functions. HTML is specific mark-up for use in displaying documents on web. XML is a standard for the creation of mark-up languages for use on the web.HTML has some limitation that includes restricting the user to a relatively small set of tags. It has limited set of tag (not extensible). Authors cannot create their own HTML tags. Another limitation of HTML is tags that control presentation are in the same file with tags that describe the document. XML overcomes the limitation of HTML and other mark-up languages, while providing capabilities that are not a part of the earlier languages. Here's a simple XML document and an HTML document shown in Table 1 that contain the same data [11].

Table 1
Sample Data Shown In XML and HTML Format

XML document	HTML document
<? xml version="1.0"	<html>
standalone="yes"?>	<h1 id="MN">State</h1>
<state stateid="MN">	<h2 id="12">City</h2>
<city cityid="12">	<dl>
<name>Johnson</name>	<dt>Name</dt>
<population>5000</population>	<dd>Johnson</dd>
</city>	<dt>Population</dt>
<city cityid="15">	<dd>5000</dd>
<name>Pineville</name>	</dl>
<population>60000</population>	<h2 id="15">City</h2>
</city>	<dl>
<city cityid="20">	<dt>Name</dt>
<name>Lake Bell</name>	<dd>Pineville</dd>
<population>20</population>	<dt>Population</dt>
</city>	<dd>60000</dd>
</state>	</dl>
	<h2 id="20">City</h2>
	<dl>
	<h2 id="20">City</h2>
	<dl>

In the XML document, the tag names convey the meaning of the data they contain. The structure of the document is easily discerned and follows a pattern. In contrast, the 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.

4. PDF

PDF, the portable document format, was invented by adobe systems in order to provide a system independent way of delivering page-based information. It is designed for brochures, magazines, forms, reports images and other materials with complex visual design, which will be printed on postscript (tm) printers. PDF file retains the exact appearance of a document, no matter what platform is used to view or print it. Anyone can view these files on their computer if they have Acrobat reader.

The format was created to remove machine and platform dependence for the documents, and its goals include design fidelity and typographic control. It was never designed for interactive online reading. However, many word processors, page layout and other programs can create PDF files easily; so many sites are now serving them online. PDF file are created by printing to a PDF drivers or by "distilling" a postscript file.

5. IMAGE

Image is the application of digital technology to the management of information existing in non-digital format such as paper, photographs, microforms, and voice. A digital image is an image that a computer can store, read, and display. It is composed of a set of pixels arranged according to a predefined ratio of columns and rows. Each pixel presents a portion of the image in a particular colour or shade of grey (Getty Research Institute, 2000).

"Image" is the graphical representations of real-world objects. Image can be representing through the development of photography, video, computer data. Still and moving images can now be stored and transmitted in digital form. This allows images to be stored, transmitted and manipulated by computer in different type of formats. The most commonly used images file formats are

Tagged Image File (TIFF) File extension *.tif

Graphics Interface Format (GIF) File extension *.gif,

Joint Photographic Expert Group (JPEG) File extension *.jpg PC Paintbrush Format (PCX) File extension *.pcx,

Standard Windows Bitmap BMP) File extension *.bmp, Portable Network Graphics (PNG) File extension *.png, Photoshop images (PSD) File extension *.psd,

Macintosh format (PICT) File extension *.pic or *.pct,

Pixar Image Computers (PIXAR) File extension *.pxr, SciTech continuous tone (SCITEX CT) File extension *.pxr, True vision video board (TARGA) File extension *.png,

Raw format (RAW) File extension *.raw.

All image filed have two parts. The first part knows as the file header contains information about image type, colour schema and image width and height. The second part, image data contain the pixel information that actually makes up the image. Image data are often compressed in different ways to reduce file size. It is important to be aware of different file formats and compression techniques, because they effect respectively, file compatibility and information content.

6. SEARCHING PARAMETERS OF ABOVE MENTIONED FILES FORMAT

A general and extremely useful feature of digital information is the way that can search easily for specific strings, or words and phrases. In some cases, it might be possible to carry out more sophisticated searches. The World Wide Web has become such a successful channel in delivering and sharing information that people are getting used to searching the web as the first resort for information,. As the amount of data accessible via the web grows rapidly, the weakness of traditional ways of browsing and searching the web becomes more and more apparent (Laender, 2002). Browsing requires users to follow links and to read (usually) long web pages, thus making it tedious and difficult to find a particular piece of information. Keyword searching usually returns massive irrelevant information, along with some useful information hidden in the long list of search results. Even with improved search engines, such as Google, that return accurate results, a large number of web pages cannot be indexed by those engines. Therefore, users surfing the web with these traditional facilities have been facing the information overload problem; they are overloaded with too much irrelevant information. Thus authors should decide that in which format he/ she should preserve their information for pin-pointed, exhaustive and expeditious searching.

Lets us discuss the searching criteria of information preserved in HTML, XML, PDF and Image format.

(A) HTML and Searching

The most common web files type that holds Meta as well as full-body text information is HTML. The documents hold in HTML format can perform to search both Meta and full text data by using different types of text retrieval search engines, such as dtSearch Text Retrieval Engine. It also performs both Boolean or Proximity search. After a search, the user would have customizable, browser-based document sorting, document hit, and document navigation options. The retrieved documents would appear in the browser with the hits graphically marked, as well as all HTML links operational, and all embedded images intact.

But HTML doesn't give us a way to describe the contents of the text: the meaning is lost because there is no way to tag it. For instance, if you have a catalogue of hand carved doors, you probably want to talk about the size, weight, material. It would be great if your browser would sort the list in various ways or let you import the list into a database. HTML sacrifices power of ease of use and as a result there is nothing in HTML to distinguish one table or one heading from another, except for the keywords enclosed within tags.

(B) XML and Searching

The development of XML is solving the search problem of HTML. The major benefit of XML is the possibility for vastly improved web-based search. By this, they mean that instead of searching the whole text of a page, search engines could use the XML tags to specify which parts of the pages to search, as field, which should improve and provide more precise listing of the information available.

At present, most search engines pay little attention to mark-up, and focus instead on the content of the page. Consequently, results are produced mainly from the information found in the <TITLE> tags, or somewhere in the <BODY> of the document, the equivalent of full-text hits. For example, if one were to search "Mark Twain" on the Web, one could find the document shown in Figure 1 and 2 in different ways.

This page might be ranked highly by certain search engines for the following reasons:

1. It contains the exact term in the title of the document
2. The term appears early in the document
3. The term is repeated in the document

```
<HTML>
<HEAD>
<TITLE> Mark Twain
</Title>
</Head>
 <Body>
 <H1>Mark Twain<H1>
Nationality: American<P> Period:
American<P> Genere: Fiction <P>
Summary: Mark Twain was the pen name of Samuel
Clemens, an American humorist who lived from
1835 - 1910
Works:
<UI>
<LI>Adventures of Huckleberry Finn - 1884
<LI>A Connecticut Yankee in King Arthur's court - 1889
</UI>
</BODY>
<HTML>
```

Figure 1: Possible Document Result for "Mark Twain" Search on the Web

Whereas the page that actually deals with the author might reasonably mention Mark Twain once or twice, a business might repeat the name often, and thus be interpreted as more relevant. The problem here is that search tools operate without context. One could attempt to improve the search results by adding terms like "literature" or "writer", but unless the author of a page saw fit to include such terms, this strategy will have little impact on the quality of the pages retrieved.

Meta-tags could be a tremendous aid to Web search tools, included in the header of the document; one can ascribe keywords to describe the content of the document through use of the "keyword" attribute. This allows Web authors to create the context that is solely lacking in most HTML documents:

<HTML>

<HEAD>

<META NAME= "KEYWORD" CONTENT= "American Literature, Authors, Mark Twain, Works">

<TITLE>Mark Twain

</TITLE>

</HEAD>

<HTML>
<HEAD>
<TITLE>
Mark Twain Insurance Company
</TITLE>
</HEAD>
<BODY>
The Mark Twain insurance has been in business since 1956. During that time, the folks at Mark Twain have…
Call Mark Twain insurance today.
</BODY>
</HTML>

Figure 2: Another Possible Document Result for "Mark Twain" Search on the Web

(C) PDF and Searching

PDF files are hard on search engines, and HTML pages are much easier for them to deal with. While PDF, unlike HTML, contains its own built-in text search functionally, the built-in search functionality does not operate over the web. A number of third party search tools do provide web-based PDF text searching comparable to HTML searching, including such features as full display over the web retrieved PDF files with highlighted hits, all images intact.

But combining full text searching with the ability to hold multiple searchable Meta fields is a different matter. Abode includes four fields with PDF field: title, author, subject and keywords. But for many advanced web data warehousing needs, four fields are not enough. For example, an organization might want to add a department field, a project ID field, a data field etc.

PDF Web search, which uses an overlay to the PDF document format, provides for extra user-defined fields. Using a proprietary overlay to the Search Engine, it searches documents based on the characteristic of these additional fields, in combination with full text searching. In fact, it simplifies this type of combined field and non-field searching by using drop-down menus and cheek boxes to represent built-in fields, and combining that with a full text search box.

PDF Web search makes full use of the PDF file format by, for e.g. providing full support for both hidden and non-hidden document summaries. In recognition of the fact that many PDF documents are very long, it supplies instant navigation to the location of a search hit. In this way, the user would not have to download 197 pages of a 200-page document before getting the hit, but could instead immediately jump to page 197. The product also provides dynamically resort able search results and other bells and whistles, as well as support for HTML and XML. PDF file have full text search features for locating words, bookmarks, and data fields in documents. PDF file has weak searching facilities compare to both HTML and XML.

In PDF file there are some limitations for searching. These are Documents, which were scanned directly into PDF may only have the graphic portion: there may be no computer-readable text at all. These documents are not searchable.

Documents that were scanned and converted from graphic display to digital text using OCR (Optical Character Recognition) may have significant numbers of errors. In this case, many search terms will not be matched although the words were in the original printed or typed text, because they were not correctly interpreted.

Documents with multiple columns, which were converted to PDF by some layout programs, will display correctly and contain the correct digital text, but they miss the text flow: the words don't come in the correct sequence. Therefore the search engines will fail to match phrase queries because the phrases were wrapped on the next line of the column in the original, but that relationship was not stored in the PDF.

Documents generated by some applications will contain partial words due to hyphenation, incorrect coding of ligatures and extended characters (diacriticals and letters beyond the basic 26), and other unusual situations. These mangled words will not match queries, although the words were in the original text.

(D) Image File Formats and Searching

The imaging software permits instant identification and retrieval of individual documents and even of information within documents. Information created in an image format and store in a database is called digital Image database. Today, a growing number of digital image databases and libraries are available, and are providing usable and effective access to image collections. In order to access these resources, users need reliable tools to access images. Because of the huge amount of information, it is like looking for a needle in a haystack. The tool that enables users to find and locate images is an image search engine (ISE). Different Image Search Engines (ISEs) have their different features but most common search features of many ISEs are: (Hassan and Zhang, 2001)

- Key word related search: include keyword searching, which is one of the basic and most useful features in any ISEs.
- Search limitation: include the ability to limit retrieved items to a certain files format or a specific file size. The second limitation relates to the physical image limitation, such as resolution (enables users to limit their search to high, medium or low resolution for the retrieved images), orientation (allow users to control a retrieved image set to horizontal, vertical panoramic or square images), colour (user can restrict their search to colour or black and white images) and picture type (It enables users to narrow their search to photo, graphics or illustrations).
- Full text searching is possible if the documents are scanned with OCR technology Berinstein and Fieldman (1996) outlined some of the characteristics of the ideal ISE, saying that it should:

Allow keyword searching of image content, date and creator; Let users search by colour, shape and other formal attributes; Search database internal to a site Display the image as part of the search results Allow users to find the rights-holder Furnish the rights status and terms for licensing.

7. COMPARATIVE ANALYSIS OF ABOVE MENTIONED FILE FORMATS

Table 2: Comparative Analysis of different file formats that can be used in WBIS

Table 2

	Support System	Web-based information storage format			
		HTML	XML	PDF	Image Format
1	Proprietary file type	No	No	Yes	No
2	Require browser add-on for viewing	No	No	Yes	Yes
3	Supports fields along with text	Yes	Yes	Four fields	No
4	Supports nested field	No	Yes	No	No
5	Full control over image and text display in browser	No	No	Yes	Yes
6	Searching by tag names, tag attributes, data content and location within a document	No	Yes	No	No
7	More secure of document exchange (password protected)	No	No	Yes	No
8	Allows meta-language	No	Yes	No	No
9	Full browser support/end user application	Yes	No	Yes	Yes
10	Boolean and precise search	Yes	Yes	Yes	No

8. CONCLUSION AND FUTURE SCOPE

In summary, there are different types of theories available for WBIS. It is critical for today's researcher to decipher an appropriate and applicable theory from different theories available in the IS domain.

The researcher uses the ToDA (Theory of Deferred Action) as informing practice to improve the rational development of WBIS for an emergent higher education organization. Weber (2003) makes a good argument for the need to develop our "own" theory that characterizes our field. Though Gregor (2006) appears to label the IS community as: still struggling to identify strong theories that are unique to IS. Conducting this research brings qualitative refutation to the ToDA with its potential to become a reference theory for the IS discipline and

organizational studies. Adversely, for ToDA to become a reference theory, scholars argue that its refutation needs both qualitative and quantitative testing in order to generate an applicable theory (Nagel, 1961).

Further action research is currently being undertaken to investigate how the Kadar Matrix can improve the speed of time-to-market. This will be done through monitoring (data-gathering) its effectiveness within different WBIS development projects. This aligned with ToDA can add more rigors to its actual effectiveness in actuality.

REFERENCES

[1] Ramrattan, M., and Patel, N.V. (2009), "Web-Based Information System Development and Organisational Change: The Need for Emergent Development Tools, European and Mediterranean Conference on Information Systems Crowne Plaza Hotel, Izmir, 13-14 July.

[2] Patel, N.V, and Hackney, R. (2008), "Designing Information Systems Requirements in Context: Insights from the Theory of Deferred Action". European and Mediterranean Conference on Information Systems, May 25-26.

[3] AIS (2009): Available at: http://www.fsc.yorku.ca/york/istheory/wiki/index.php/Main_Page [Accessed 9th May 2009]

[4] Berinstein, P. and Field, S (1996), Finding Images Online: Online user's Guide to Search for Images in the Cyberspace, Pemberton Press, Wilton, CT.

[5] Bourret, R (1998), "Declaring Elements and Attributes in an XML DTD", the Database Research Group at Die Technishe Universitat Von Darstadt, 3 March 1999. Available: http://www.informatik.tu-darmstadt.de/DVS1/staff/bourret/XML/Xmldtd.html.

[6] Bradley, Neil (2000), The XML Companion, Pearson Education Ltd., Harlow, England

[7] Culshaw, Stuart: http://xml.coverpages.org/culshaw SunserverXML.html

[8] Getty Research Institute (2000), "Introduction to Imaging. http://www.getty.edu/gri/standard/introimages/index.html

[9] Goldfarb, Charles F and Prescod, Paul (2001), The XML handbook; Addison Wesley Longman (Singapur) Pvt Ltd, Delhi, pp. 20.

[10] Hassan, Ibrahim and Zhang, Jin. (2001), "Image Search Engine Feature Analysis", Online Information Review, 25, No. 2, pp. 108-114.

[11] IBM WebSphere Application Server 2.0. http://www-306.ibm.com/software/webservers/appserv/doc/ v20dcstd/doc/whatis/icxml4j.html#xmlhtml.

[12] Laender, A H F; Ribeiro-Neto, B A and Silva, A S D (2002), "DEByE-Data Extraction by Example", Data and Knowledge Engineering, 40 (2), 121-154.

[13] Pietromonaco, P. (2002, August), "The Magic of HTML", Poptronics, 3(8), 16.