WEB-BASED INFORMATION MANAGEMENT : HTML, XML, PDF AND IMAGE FORMAT : AN ANALYSIS IN COMPARISON

Nihar K Patra Ashis K Pani Rajendra K Thaty

Abstract

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 everincreasing 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 behavior must also be taken into account. Appearance and behavior 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 is 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.

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 Markup Language (HTML).

In 1989, a researcher named Tim Berners Lee proposed that information could be shared within the CERN European Nuclear Research Facility using hyperlinked text documents. He was advised to use an

 © INFLIBNET Centre, Ahmedabad

SGML-ish syntax by a colleague named Anders Berglund, an early adopter of the new SGML standard. They started from a simple example document type in the SGML standard and developed a hypertext version called Hyper Text Markup Language (HTML). (Goldfarb, 2001)

HTML is just one of the many thousands of different document types that have been created using SGML. It was designed specifically to enable documents to be published on the World Wide Web (WWW). It defines a fixed set of document elements with markup that describe simple (i.e. not highly structured) documents containing headings, paragraphs, list, illustrations and so on. Over the last ten years, HTML has proved extremely successful as a simple means of publishing information in a user-friendly hypertext format. It has also been adopted to incorporate the presentation of graphical and audio material.

An HTML file consists of text as well as tags, which tell the browser how to format it. An HTML document consists of two major sections: the head and the body. The head contains details about the document, and body consists of any number of elements such as text, hypertext, tables and object references. The tags describe how the images and texts are going to appear on your site (Pietromonaco, 2002)

The major contribution of HTML, in light of the rapidly increasing adoption of intranets for producing and managing corporate information and documentation, has been to educate a very large audience as to the main advantages of distributed information system. These advantages include the ability to exchange information between different computer systems and application through the use of standard formats and protocols, and the power of hypertext to organize a set of documents to be searched, access and consulted interactively. (Culshaw)

HTML is probably the most portable markup language in the world and is becoming the de fecto standard for transmitting information between people.

2.1 Leverages

- It is simple and a open standard
- It is fairly easy to learn
- It is good at presenting text and graphic in a reasonably decent layout
- Its files are tiny.
- It allows the view of information online through the use of interactive form
- Its browsers are cheap or free, very powerful; with a combination of third party add-ins and server-side content support, a vast range of information is being delivered through HTML language
- HTML document browser interfaces are easy to build into existing products because of the simplicity of HTML
- HTML pages allow link to any other publicly accessible page simply by entering the address
- There are some specialized structures in HTML, but they are mostly used to effect a certain formatting look
- It permits a variety of other enhancement other than internal and external hyperlinks
- It is interactive

2.2 Perils

- It has a limited set of tags (not extensible) and very loose document structure rules, it offers only one way of describing documents
- It provides linking capabilities, but the linking is rudimentary; it is only a one-to-one link, and requires an anchor on the target end in order to access anytime within the document
- Content tagging is very limited: the context and formatting are inseparable tied together
- The visual presentation is dependent on the setting of browser. Thus one can never be certain that the reader will be shown precisely what the author of the HTML intended
- Only code for display, not document structure, Semitics or content
- Not extensible – can not customize
    - Cannot accommodate special needs (e.g. mathematics, chemical formulas)
    - Proprietary, vendor – specific tags to extend capabilities
- Hard copy is less compact and harder to read
- Design for reading on the screen
- Doesn't support math symbols, except as auxiliary bit-mapped files
- Doesn't support scientific notation
- Limited, predefined data structures
- No formal validation
- Trades power for ease of use
- Good for simple applications only
- Hard crafted – links, navigation, indexing
- Concentrates on form, not substance

3. XML

Extensible Markup 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) (Thomas). 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, that can be used to define an infinite number of customized markup 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 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 markup 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 markup languages. In simple terms, a document markup language is a set of element (frequently called tag) that has one or more of the following functions

- Describes the structure of the document
- Describes the content of the document
- Controls how the document is presented to the user

XML is not a set of tag itself: it provides a standard system for browser and other applications to recognize the data in a tag. Unlike proprietary formats, XML format are open to all. Consumer will benefit as we will have more control on our data, there will be many more useful little programs because its easy to read the files, and we will be able to use the data even when the original application are long gone. When applied to the web, it makes information interchange much richer and more interesting: Tim Berner-Lee calls this the *Semantic web*. (http://www.searchtools.com/index.html)

The design goals of XML are (BradLey, 2002)

- XML shall be straightforwardly usable over the Internet
- XML shall be supporting a wide variety of applications
- XML shall be compatible with SGML
- It shall be easy to write programs which process XML documents
- The number of optional features in XML is to be kept to the absolute minimum, ideally zero
- XML documents should be human-readable and reasonably clear
- The XML design should be formal and concise
- The XML design should be prepared quickly
- XML documents shall be easy to create
- Terseness in XML markup is of minimum importance

It will be more apparent, if it is compared with HTML.

HTML is specific markup for use in displaying documents on web. XML is a standard for the creation of markup 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 markup languages, while providing capabilities that are not a part of the earlier languages. Here's a simple XML document and an HTML document that contain the same data (http://www-306.ibm.com/software/webservers/appserv/doc/v20dcstd/doc/whatis/icxml4j.html#xmlhtml)

XML document

```
<?xml version=”1.0" standalone=”yes” ?>
<state stateid=”MN”>
<city cityid=”12">
  <name>Johnson</name>
  <population>5000</population>
</city>
<city cityid=”15">
  <name>Pineville</name>
  <population>60000</population>
</city>
<city cityid=”20">
  <name>Lake Bell</name>
  <population>20</population>
</city>
</state>
```

HTML document

```
<html>
<h1 id=”MN”>State</h1>
<h2 id=”12">City</h2>
<dl>
<dt>Name</dt>
<dd>Johnson</dd>
<dt>Population</dt>
<dd>5000</dd>
</dl>
<h2 id=”15">City</h2>
<dl>
<dt>Name</dt>
<dd>Pineville</dd>
  <dt>Population</dt>
<dd>60000</dd>
</dl>
<h2 id=”20">City</h2>
<dl>
  <dt>Name</dt>
  <dd>Lake Bell</dd>
  <dt>Population</dt>
  <dd>20</dd>
</dl>
</html>
```

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.

3.1 Leverages

- Provides more accurate description of document content by enabling an extensible tag set. XML implements can define their own tag sets to describe document contents, authors are not restricted to a limited set of tag defined by proprietary vendors.
- Enables validating document contents against a standardized grammar. The Document Type Definition (DTD) is an example of such a grammar. The grammar describes the valid tags, attributes (characteristic of tags, such as identifiers), and other content for the XML document.
- Makes it easier to exchange documents among users and applications. XML is best format for source documents, because it enables delivering content in the most appropriate output format (such as HTML, portable document format, & post script) and format for applications like CEDI electronic data interchange)
- Supports advanced searching. Searching by tag names, tag attributes, data content, and location within a document are other search strategies that XML documents makes easier to implement
- Complex relationship like trees and inheritance can be communicated.
- Content can be presented easily to different users in different forms e.g. an auto parts catalogue can be presented to a shopper as a view that includes the prices, descriptions, and order numbers

for parts. The catalogue view for the auto mechanic could include the information available to shoppers plus schematic that shows the position of the installed part. The manufacturer's view could include information about subcomponents and materials.

- Improves user response, network load, and server load. XML implementations can have the web server send on XML document and its associated XSL style sheets to the client once.
- Supports Unicode: the advantages of XML include greatly improved hypertext linking capabilities and provision for multilingual document encoding through build-in support for the UNICODE standard
- Supports advanced linking among document
- Allows single document to be used in many ways

3.2 Perils

- More difficult, demanding, and precise than HTML
- Still experimental / not solidified
- Lack of browser support / end user application. There are no XML browser on the market yet (although the latest version of IE does a pretty good job of incorporating XSL and XML documents provided HTML is the output) or IE 5 and Netscape 5 are expected to fully support of XML
- Standards or protocols that are current under discussion or development include: Microsoft's channel definition format, an XML-based metadata specification for publishing applications; MathMl, an XML application for describing the structure and content of math expressions; and even an XML interface for accessing information in data base.

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. Any one 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.

4.1 Leverages

- Provides electronic pages with impressive page fidelity. Type, graphics, and color are all reproduced as they are on paper
- PDF files are cheap to create and are used by many companies to deliver page formatted information without the high cost of postage
- Solves file sharing problem between platforms
- Hot links and other electronic object types, like movies and sounds, can be added to a PDF file

- Since the end user gets something that looks much like paper, training costs are low
- More secure document exchange: with the Adobe acrobat 7.0 software, PDF files can be password protected to prevent unauthorized viewing and altering. It controls whether the reader is permitted to cut/paste or make hard copy printout. It controls whether others are permitted to modify the document.
- Adobe files have full text search features for locating words, bookmarks and data fields in documents
- Share documents with any one adobe PDF documents can be shared views and printed by any one, on any system, using free Adobe Reader ® software regardless of the operating system, original application, or fonts
- PDF has facilities for web integration and delivering, including hyperlinks and forms
- Extended language support means that one can view, search, and print PDF documents that contain Cyrillic, Central and Eastern European, and Asian text.
- PDF documents have automatically generated table of contents, thumbnails and indexes.
- PDF has article threading. Scanning a Journal can leave it exactly same as the original layout. The viewer program will automatically guide the reader through the disconnected pieces of an article.
- Easier to prepare
- Email-friendly
- Perfect for forms

4.2 Perils

- PDF files are not nearly as flexible as other electronic formats because the main goal is to recreate a paper page, and not to provide ways of delivering intelligent document structure to a user
- There is limited support for searching, although Adobe has products that can index many different PDF files for cross document searching and navigating
- Problems of reading PDF files online. Harder to read unless printed
- Copying of table, image, and text simultaneously is very difficult and selection cannot cross page breaks
- The Adobe Acrobat Reader software must be downloaded to view the PDF document and Acrobat Program is needed to prepare PDF document
- Requires the Adobe Acrobat program to create or edit the PDF document
- Have to be fully downloaded before you can read them
- Not indexed by most search engines
- Don’t do interaction very well

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 color or shade of grey (Getty Research Institute, 2000).

“Image” as 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 **T**agged **I**mage **F**ile (TIFF) *File extension* *.tif Compuserve **G**raphics **I**nterface **F**ormat (GIF) *File extension* *.gif, **J**oint **P**hotographics **E**xpert **G**roup (JPEG) *File extension* *.jpg **PC** Paintbrush Format (PCX) *File extension* *.pcx, Standard **W**indows **B**itmap BMP) *File extension* *.bmp, **P**ortable **N**etwork **G**raphics (PNG) *File extension* *.png, **P**hoto**S**hop 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, Truevision 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, color 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.

5.1 Leverages

- No loss of image data
- Free exchange between application and computer platforms
- Preserves high-quality color separations, useful for producing very high quality prints

5.2 Perils

- Text captured in a image cannot be searched or otherwise accessed by electronic means
- Image files are typically much larger than files containing text, and are thus more expensive to store and slower in distribution using computers
- Text cannot be modified easily

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.

6.1 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.

6.2 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 markup, 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 I and II in different ways.

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

- It contains the exact term in the title of the document
- The term appears early in the document
- The term is repeated in the document

Figure I Possible document result for "Mark Twain" search on the Web

```
<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 II Another possible document result for "Mark Twain" search on the Web

<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>

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>

Since XML allows for the creation of tag sets that are content savvy, an XML data structure can serve as a road map to information, as with the example shown in Figure III.

Figure III: Example of XML data structure serving as road map to information

```
<?xml version=”1.0" encoding=”UTF-8?>
<DOCTYPE AUTHOR SYSTEM “author.dtd>
<AUTHOR>
        <NAME>Mark Twain</NAME>
        <NATIONALITY>American</NATIONALITY>
        <PERIOD>19th Century</PERIOD>
        <GENRE>Fiction</GENRE>
        <WORK>
          <TITLE>Adventures of Huckleberry Finn</TITLE>
          <YEARPUBLISHED>1884</YEARPUBLISHED>
          </WORK>
          <WORK>
                    <TITLE>A Connecticut Yankee in
          King Arthur’s Court</TITLE>
          </WORK>
        </AUTHOR>
```

A search engine would no longer be looking solely at the information inside the tags, but at the tags themselves. The tags in Figure III create a logical hierarchy that would situate search terms within the necessary context, and thus eliminate the irrelevant types of hits so common to keyword searching in a Web of full-text documents. Thus an XML savvy search engine, which would rank markup hits much higher than content hits, would be able to distinguish quite easily between Mark Twain, the author and the Mark Twain, the Insurance Company.

XML is able to accomplish these structuring tasks through the language’s ability to associate an XML document with a document type definition (DTD). The DTD is where the structured tags are actually declared and attributed to certain values. DTDs can be included at the beginning of the XML document, or maintained as an external file. The external DTD, author.dtd, is declared in Figure IV.

Figure IV - External DTD (document type definition)

```
<?xml version=”1.0" encoding+”UTF-8"?>
<!DOCTYPE simple [
<!ELEMENT AUTHOR (NAME,NATIONALITY,PERIOD | GENRE+,WORK+)>
<!ELEMENT NAME (#PCDATA)>
<!ELEMENT NATIONALITY (#PCDATA)>
<!ELEMENT PERIOD (#PCDATA)>
<!ELEMENT GENRE (#PCDATA)>
<!ELEMENT WORK (TITLE, YEARPUBLISHED)>
        <!ELEMENT TITLE (#PCDATA)>
        <!ELEMENT YEARPUBLISHED (#PCDATA)>
]>
```

DTDs are optional in the XML 1.0 specification. Documents that have an associated

DTD are said to be “valid” XML. Documents that conform to the rules of XML, but do not have a DTD are said to be “well-formed”. Well-formed XML would be an attractive option for Web authors wishing to convert HTML to XML, without going to the time and expense of creating a DTD. However, DTDs provide

the key to much of the promise of XML, since they give the end user an efficient means of associating search terms with what is in effect a list of key terms. In Figure IV, an ELEMENT has been declared for each item pulled from Figure III. ELEMENT AUTHOR is actually a group element composed of the name, nationality, period, genre, and work, and subelements are then defined further through the list. The "+" symbol following GENRE and WORK is similar to the truncation with which librarians are familiar, declaring these group elements as having one or more genre types or that the author has more than one published work. Notice also the "|" symbol found between PERIOD and GENRE which also relates to the Boolean logic used by those who search databases. In this element group, the DTD will associate the XML document with either the period or genre in which it was written, depending on whether both elements appear in the document. PCDATA stands for Parsed Character DATA and basically implies that the information found within the element will be character data as opposed to numerical data (Bourret, 1998)

While standard tag sets are derived, one can envision the potential for precise searching. As seen in Figure III, and implicitly through the element declarations in Figure IV, online users could limit searches by author, title, the time period in which a work was written, and, perhaps serendipitously, by Library of Congress classification schemes. Considering the plethora of different types of files found on the Web, XML even allows for searching by document extension.

XML enabled browser could display the search results of a query alphabetically, chronologically, by subject sets, or even by file type and language. Basically, XML allows users to search for and manipulate data found on Web sites to suit their own needs.

Thus, due to semantic nature of XML elements, search can be restricted to elements, thereby pinpointing searches. XML searching is an acknowledged search problem and we are only beginning to explore the exact benefits of searching XML files. There are a variety of searches that can be performed.

1. One can search for one or more keywords inside of one or more elements. The keywords can be treated as a conjunct by including the keyword "and" in the search term. Otherwise, the keywords are treated as disjunct.
2. One can search for keywords in entire documents
3. One can browse pages by element

7. 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 Websearch, which uses an overlay to the PDF document format, provides for extra user-defined fields. Using a proprietary overlay to the dtSearch 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 Websearch 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 resortable 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

8. Image files format 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), color (user can restrict their search to color 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 color, 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

9. Comparative analysis of above mentioned file formats

	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

10. References

1. Berinstein, P. and Field, S (1996), Finding images online: Online user's guide to search for images in the cyberspace, Pemberton Press, Wilton, CT.
2. 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
3. Bradley, Neil (2000), The XML companion, Pearson Education Ltd., Harlow, England
4. Culshaw, Stuart: http://xml.coverpages.org/culshawSunserverXML.html
5. Getty Research Institute (2000), Introduction to Imaging. http://www.getty.edu/gri/standard/introimages/index.html
6. Goldfarb, Charles F and Prescod, Paul (2001), The XML handbook; Addison Wesley Longman (Singapur) Pvt Ltd, Delhi, 20 p.

7. Hassan, Ibrahim and Zhang, Jin. (2001), “Image search engine feature analysis”, Online information review, Vol. 25, No. 2, pp. 108 – 114

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

9. 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.

10. Pietromonaco, P. (2002, August), “The Magic of HTML”, Poptronics, 3(8), 16.

11. Thomas, Martin. Electronics text notes – etx.xml. http://www.etext.leeds.ac.uk/cocoon/etx/lect/etx.xml?

12. XML and search. http://www.searchtools.com/index.html

About Author

Shri Nihar K Patra is a Librarian at Sir Jehangir Ghandy Library, XLRI, Jamshedpur

Dr Ashis K Pani is a Professor and Chairman of Information Systems Area, XLRI, Jamshedpur

Shri Rajendra K Thaty is a Librarian at Sambalpur University.