DYNAMIC FONT TECHNOLOGY AND SEARCH ENGINE IN NEWSPAPER CLIPPING WEB-SERVICE: AN EXPERIENCE IN CAT LIBRARY

Indu Bhushan A Bharti J K Pattnaik

Abstract

A web-based newspaper clipping service is provided by Centre for Advanced Technology (CAT) Library to it's user community. This paper describes our experience in designing a web-based newspaper clipping service on CAT Intranet. The authors describe in this paper it's coverage, digitization procedure, web enabling techniques and search engine. This paper provides a brief overview of dynamic font technology and its implementation in multilingual newspaper clipping web-service. It also discusses future plans.

Keywords: Newspaper clipping, Web-service, Search engine, Dynamic font technology, Embedded font.

1. Introduction

Newspaper gives truth without any favour. Newspaper is fast and authentic information source. Comparatively speaking the information provided by newspaper is of higher authenticity, and once it is issued, the information is recorded in a stable and unchangeable way, which gives newspaper an advantage over other media. Newspaper as a source of information has certain features like abundance, varieties of information contents that make it different from other information source such as book, periodical, radio, television and internet. The huge volume of information in newspaper is usually timely and instant, and has huge quantity of circulation.

The fields of web-based library service and digital libraries have been attracting many research efforts during the last years. Many interesting service have been started based on these techniques. The newspaper clippings web-service is a digital collection of local and national newspaper articles. Earlier news paper clipping used to be done manually. Thanks to modern technology, newspaper clipping is now less labor intensive. Print media is useful for research needs but many organization and individuals are turning to online newspaper clipping services and some are organization do this by their library.

Traditionally CAT Library has been disseminating the news related to CAT and Department of Atomic Energy (DAE) after searching the English and Hindi news papers subscribed by the library. We used to put the photocopy of the news on notice board. We also used to send a copy of the news published in local news paper to the publication division of DAE as a part of CAT Library routine work. With emergence of new technology like e-mail, networking, digitization, CAT Library started digitizing the newspaper clipping and putting them on the intranet and sending e-copies of the newspaper clippings to DAE by e-mail. This made the work easier and faster. With the increase of number of clippings it was also necessary to attach a search engine for retrieval purpose. Due to the complicacy of loading fonts for Hindi clipping, we have started using dynamic font technology.

2. Coverage

We scan news related to DAE and CAT from the following six English daily newspapers and four Hindi daily newspapers regularly (see Annex 3a & 3b).

1. Free Press
2. Hindustan Times
3. The Economic Times
4. The Hindu
5. The Indian Express
6. The Times of India
7. pkSFkk lalkj
8. ubZnqfu;k
9. uo Hkkjr
10. nSfud HkkLdj

CAT library subscribes these above newspapers. Division on news is primarily based on newspaper languages i.e. Hindi & English, and then further divided in two categories, 1. News of DAE and 2. News of CAT. The link to the news paper clippings has been provided from our library homepage (Annex - 1).Further news are listed as current and previous month.

3. Digitization and web-enabling

We are using HP Scanjet 7400c scanner for scanning the newspaper. It has up to 2400 dpi with true 48-bit color using HP's exclusive scanning technology. It is capable of scanning of negatives, slides and transparencies. It can be shared between users using precision LAN software. Scanning software includes fully integrated Optical Character Recognition (OCR) capability.

3.1 Type of Image files used and their naming

HP precision has capability to save file in desired format (jpg, tiff, bmp etc) but we keep the file in 'jpg' format on the net and 'tiff' files for preservation as 'tiff' format has long durability but takes more space. We also write news paper name and date on the image. A scanned page of a clipping has been shown in Annex – 5.

3.2 Editing and Web-enabling

After scanning it has to be edited as some paper has bad print quality or some spot. It can be better by changing brightness, contrast etc. After scanning files are kept in their respected folder. Permission is given to file and news are added in news list with proper hyperlink.

3.3 Folder and files management

News are divided in two categories and further two sub divisions same of the same is done in folder management.

File are named as dd + mm + year + _short name of paper + file extension. For example, a news published in Dainik bhaskar on 12/11/05 will be named as 12nov05_db.jpg where db means Dainik Bhaskar

http://library.cat.ernet.in/homepages/news/DAE/hind 1
/eng 2 for DAE news

http://library.cat.ernet.in/homepages/news/CAT/hind 3
/eng 4 for CAT news

http://library.cat.ernet.in/homepages/news/CAT/font 5 for .eot font files

These folder and files are stored on CAT Library Linux server.

3.4 Mailing of the news

We send the scanned news to CAT director and DAE publication division by e-mail.

3.5 Hardware and Software used

Library web server with Linux OS, Desktop PC with WIN 2000 OS, HP 7400c Scanner, Microsoft Front page, HP Precision Scanning software, Web Embedding fonts Tool and Search engine.

4. Implementation of Dynamic Font Technology

4.1 Dynamic or Embedded Fonts

Embedding a font is the technique of bundling a document and the fonts it contains into a file for transmission to another computer. Embedding a font guarantees that a font specified in a transmitted file will be present on the computer which receives the file. Not all fonts can be moved from computer to computer, however, since most fonts are licensed to only one computer at a time. Only TrueType and Open Type fonts can be embedded.

Applications should embed a font in a document only when requested by a user. An application cannot be distributed along with documents that contain embedded fonts, nor can an application itself contain an embedded font. Whenever an application distributes a font, in any format, the proprietary rights of the owner of the font must be acknowledged.

It may be a violation of a font vendor's proprietary rights or user license agreement to embed any fonts where embedding is not permitted or to fail to observe the following guidelines on embedding fonts. A font's license may give only read-write permission for a font to be installed and used on the destination computer. Or the license may give read-only permission. Read-only permission allows a document to be viewed and printed (but not modified) by the destination computer, documents with read-only embedded fonts are themselves read-only.

Read-only embedded fonts may not be unbundled from the document and installed on the destination computer.

4.2 Advantages of embedded fonts are

- Content is easy to edit and dynamically update. It can be copied and pasted.
- Search engines can spider it.
- The browser can enlarge it.
- It's fast - this whole font is only 8KB and it's stored in cache.
- No need for alt tags.
- The advantage of this technology is that the site can be viewed without downloading any fonts.

4.3 Tools and Procedure of implanting dynamic font- overview of WEFT

Many tools are available for using dynamic font technology but we are using Microsoft WEFT 3, which is freely available over Microsoft website.

The Web Embedding Fonts Tool 'WEFT', lets Web authors create 'font objects' that are linked to their Web pages so that when an Internet Explorer user views the pages they'll see them displayed in the font style contained within the font object.

WEFT software has very simple procedure to implement embedded font technology. In the very first version of Internet Explorer released in 1995, Microsoft pioneered font support was done by including support for the FONT FACE tag. IE versions 3, 4 and 5 built on this support by including support for parts of the W3C's Cascading Style Sheet 'CSS' standard. CSS gives Web site designers greater control over font specification and substitution.

Although these methods of specifying fonts within Web pages provide Web page designers with a high level of typographic control, they rely on the specified fonts being installed on a reader's computer. WEFT overcome this limitation. WEFT is a stand alone tool that lets Web page designers create embedded fonts that can be used when displaying their Web pages.

4.4 Sending fonts with your page or what is font embedding for Web pages

Dynamic font technology allows you to send fonts with your Web page. These fonts are displayed by both Internet Explorer 4 and Navigator 4 and above versions. The Web Embedding Fonts Tool WEFT', lets Web authors create 'font objects' that are linked to their Web pages so that when an Internet Explorer user views the pages they'll see them displayed in the font style contained within the font object.

4.5 Downloading WEFT its System requirements and Installation

Complete version of WEFT 3.2 that includes database components in a single file WEFTIII2b1.exe - 9.26 MB from location - http://download.microsoft.com/ can be downloaded.

WEFT has been tested on Windows 95, 98, NT 4.0 and Windows 2000. Full system requirements are listed below.

- *Disk space*: 20MB free space.
- *Web browser*: Internet Explorer 4 and newer .
- *Platform*: Windows 98, Windows Me, Windows 2000 or Windows XP.
- For installation on WIN 2K and XP user should administrator or power user.

4.6 Steps

1. Create html file using whichever font we want to use for dynamic font.

2. To start WEFT double click on the WEFT icon or select 'Microsoft WEFT' from the 'Open Type Tools' section of the Windows 'Start', 'Programs' menu. When we run the tool for the first time WEFT 3 will generate a database containing information about the fonts installed on computer and it will ask for creator mail id and name. These information will be inserted in webPage.

Weft.exe

Figure-1 Figure-2

3. Go to file menu click on new then click on next, give the name and email if asked.

4. Browse your html file, for which dynamic font have to be created and add the file then click next (Fig.-3).

5. Now next click will analyze the font in system and display the list of font and font problems. Click on next for next menu.

6. Now give the detail where dynamic font (.eot files) has to be created and url of web directory from which .eot font files have to be used (Fig. -3).

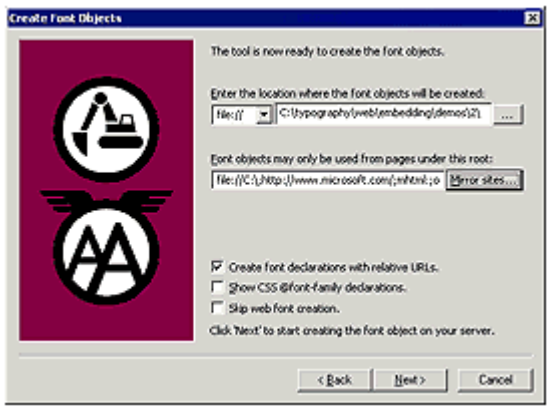

Figure-3

7. On clicking on next button it will create the font files in the directory we have given. It also write the tag in html file for picking dynamic font from the url we have given in procedure. Now tick on upload the page on server and finish.

8. Now upload html file on the web and upload .eot files in the wed directory as given in procedure.

5. Implementing Search facility

A search facility has been made available for news paper clippings. Web Based “Simple Search Engine” Software utility freely available on the Internet, has been downloaded and used for enabling search. It is a CGI PERL based programme. It can be downloaded it from URL: http://www.cnctek.com/bizdb-free-search-engine. A sample of search is shown in fig 4 & 5.

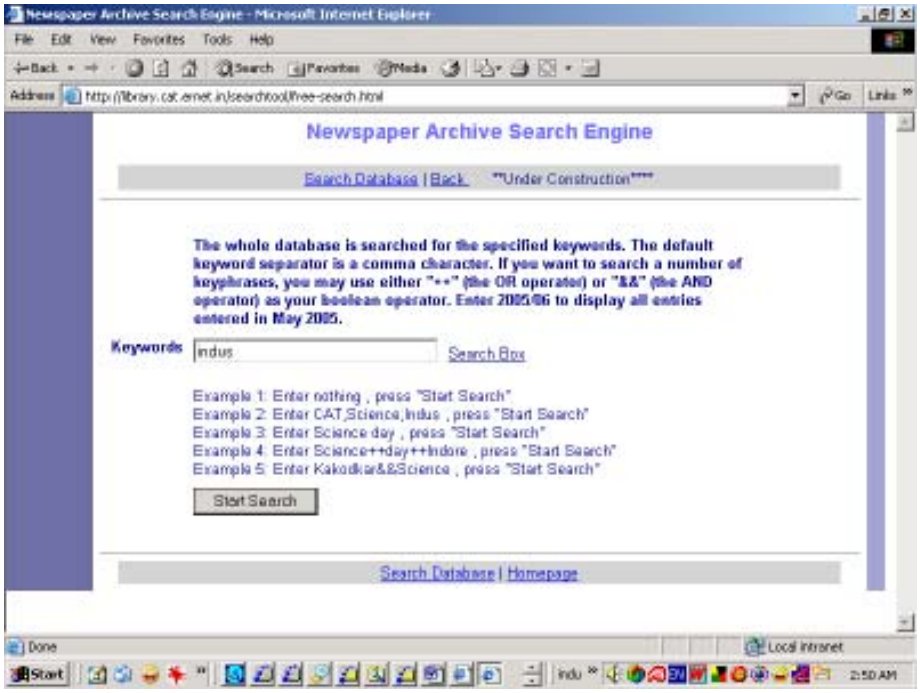

Figure -4

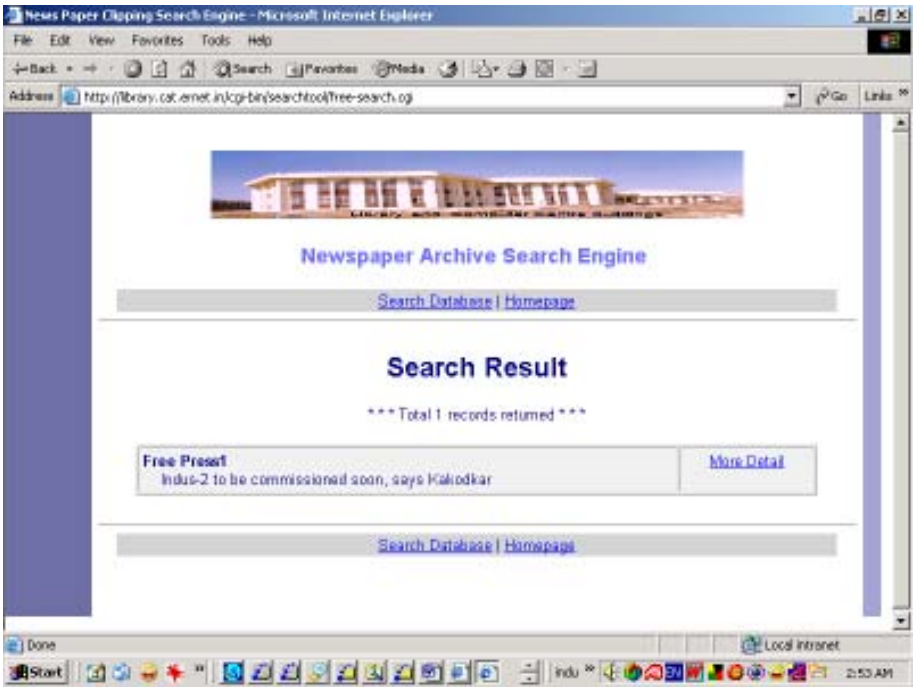

Figure -5

5.1 Step of implementing the software

We downloaded the zip file of software from the site and unzipped it. After unzipping we got following files:

1. **free-search.cgi** - CGI script used to Search and Display multiple database records at one time.
2. **free-search.html** - With sample Search DB records Html template. We have used this template to specify search criteria (keywords) for the free-search.cgi program.
3. **free-search-summary.html** - The sample Display DB Records Html template. We have used this template to define which and how DB fields are to be displayed in the search report. Fields belonging to the same record are displayed in a group. This is repeated for each record until all records or a group of records are displayed.
4. **free-search-detail.html** - This sample Html template is used to display details of a given record. This, in conjunction with the “free-search-summary.html” template, canbe used to construct a simple search engine.
5. **ReadMe.txt** - This file describes how to configure search engine to use.
6. **sampledb.txt** - Sample flat ASCII database file.

7 **testing.cgi** - Execute this test script to test cgi-bin directory and it’s working properly. This script will also help to find out the absolute pathname of your Website’s home directory.

We have made the changes in the above files and uploaded these files to our website as instructed in readme.txt file. We have changed the url, gave url of our site and image files directory, changed html templates according to our newspaper bibliographical field. Created and updated sampledb.txt.

7. Future plan of our service

1. 'PDF' format with OCR will be used to enhance the searches with in the file.
2. Simple entry form will be used for giving bibliographical detail that will be saved in backend database and it will be used both news listing at front end and searching facility so that database flat file .txt and news list html files which are to be uploaded frequently, need not to be made.
3. Scientific news of interest to our scientists and engineers will be added.
4. Yearly CDs of clipping will be kept as archrivals.

8. Conclusion

Librarians used to provide newspaper clipping services for their clients with scissors and paste two decades ago. Now with computer, network and digitized newspaper database, they are able to fulfill that task only with few mouse clicking. Librarians today can use vaster and various resources to do their information reference, to provide service more timely and effectively. They can also use multiple ways to deliver their service and communicate with their clients.

Web-based service has reduced the work of Library staffs. It has reduced the time lag for reaching its reader. It is now accessible at their desktops on a click of mouse. Search engine has made the process of retrieval easy. With the dynamic font technology newspaper clippings of Hindi newspaper has also become easy. As most libraries provides newspaper clipping service to its user community, we suggest a web-based digitized newspaper clippings service.

9. References

1. S. K. Sonkar …[et al]., Application of Greenstone Digital library (GSDL) software in Newspaper clipping, DESIDOC Bulletin of Information Technology.(Vol.32, No.3pp.).

2. Arun Maurya… [et al]., Creating website in Hindi, DESIDOC Bulletin of Information Technology. Vol.22, No.2

3. J.K. Pattnaik, R. Dighe, P. Rajendiran, A. Bharti, A. Deshpande, Indu Bhusan, Digital Collection development programme in CAT Library: an experience, International Conference on Information Management(ICIM)-2005, Mumbai, 22-25 Feb,2005 .

4. URL: http://www.ifla.org/VII/s39/conf/Shanghai.pdf

5. URL:http://www.ifla.org/VII/s39/conf/Njnormal1.pdf

6. URL:http://www.cyber-north.com/business/service.htm

7. URL:http://www.microsoft.com/typography/web/embedding/weft3/default.htm

About Authors

Indu Bhushan received B. Sc. in 1997 from Govt. Degree College, Lalitpur (UP), BLIS and MLIS from Jiwaji University, Gwalior and PGDCA from RCC, Delhi. He Worked with IGNOU (Delhi), Amity School of Engineering and Technology (Delhi) and Saha Institute of Nuclear Physics (Kolkata). Since 2002, he has been with Library, Centre for Advanced Technology (Indore) and involved in Website development and Web-based library services in library.

Anuranjana Bharti graduated in science from Panjab University, Chandigarh. She received B.L.I.Sc and M.L.I.Sc from Guru Nanak Dev University, Amritsar. Presently she is working in Centre for Advanced Technology(CAT) Library and is involved in various library technical services. She has also been involved in web-based digitized press release service

J. K. Pattnaik was born in Orissa, in 1966. He received B.Sc from Utkal University in 1987, M.Sc in Chemistry from Agra University in 1989, B.L.I.Sc and M.L.I.Sc from Delhi University in 1990 & 1991 respectively. Since 1992 he has been working in Centre for Advanced Technology(CAT) Library, Indore. Presently he is working as Officer-in-Charge of CAT Library