Case study

NewsNIC: a Web-based, full-text news clipping service from the National Informatics Centre Library in India

Ram Kumar Matoria
P.K. Upadhyay and
Arvind Mishra

The authors

Ram Kumar Matoria is a Scientist-C and Officer-in-Charge, P.K. Upadhyay is a Scientist-C and Arvind Mishra is a Scientific Officer 'SB', all at the Library and Information Services Division of the National Informatics Centre of the Ministry of Communications and Information Technology, New Delhi, India. E-mail: rkmatoria@hub.nic.in

Keywords

Worldwide web, Libraries, Library automation, India

Abstract

Newspapers constitute an important source of information as they contain the latest information in the form of news with, often daily, updating. Being aware of their importance, libraries have been providing news clipping services in various ways to their users since newspapers were invented. Initially, manual methods of marking, cutting and pasting the useful news items were used, but today's libraries have automated this service by using information technology. This certainly has improved the access, delivery, and searching of news clippings in a Web-based network environment for remote users. This case study discusses the needs, objectives and advantages of NewsNIC, a Web-based full-text news clippings project implemented by the library of the National Informatics Centre, Ministry of Communications and Information Technology, Government of India. The authors discuss the design aspects, various systems components, Web technology and database tools used as back-end solutions, and the use of the Microsoft MS SQL Server features and active server pages technology.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0033-0337.htm

Program: electronic library and information systems
Volume 37 Number 3 2003 pp. 181-189
© MCB UP Limited ISSN 0033-0337
DOI 10.1108100330330310484387

Introduction

Newspapers constitute an important source of information as they contain the latest information in the form of news with, often daily, updating. The information published in newspapers covers every aspect of the socio-techno-economic scenario prevalent in individual countries and around the world. Being aware of the importance of information published as news in newspapers, libraries have been providing news clipping services in various ways since newspapers were invented. Initially, libraries used manual methods to mark, cut, and paste the news items from newspapers and then circulate these among key personnel in an organisation.

With the advent of IT during the last few decades, libraries have started to automate this service by designing bibliographic databases of news clippings and providing online access for the users. This certainly has facilitated wide circulation of the news clippings and better searching of the archives. However, in such systems the news clippings themselves were often delivered in print format. More recently, these bibliographic databases have been supplemented by the inclusion of the news clippings as scanned images. A number of libraries and agencies have implemented this kind of news clipping imaging system (Mühlberger, 1999; ePRO System Limited, 2002; Singh *et al.*, *2000;* Pownikar *et al.*, *2002)*. Such systems have eliminated the delivery of news items in print and instead the news items are provided in digital form as scanned images. However, such systems may provide for limited searching of the archive file of news images.

With the introduction of Web technology, the availability of better scan/OCR (optical character reading) software and the advent of relational database management systems, some libraries have opted to implement a Web-based full-text news clipping system. In such a system, the news clippings are converted to text/HTML format and stored in the appropriate field in a database along with other fields such as title, author, date, paper name, etc. Such full-text databases of news clippings are then accessed by library users over the Web in a network environment. This certainly has improved the access to news clippings over the Web and archives searching from the full-text body of the news items.

NewsNIC project

The access, delivery and retrieval of a news clippings service in the traditional manual manner is a time-consuming job, especially in a large organisation such as ours where library users are posted all over the country. Therefore, it was decided to undertake the NewsNIC project to develop a Web-based, full-text news clipping system in the library for better and quicker delivery of useful news to remote users over the parent network NICNET. The NICNET is a nationwide satellite-based communications network set up by our parent organisation, the NIC, serving the Government of India (NICNET, 2002).

The main objectives of the NewsNIC project are:

- to eradicate the traditional manual method of the news clipping service and thus, save manpower time and effort;
- to provide full-text of news items in digital form with better archive retrieval techniques;
- to make use of IT tools in the library and thus to set an ideal example of a Web-based full-text news clipping service; and
- to implement such a system in other government libraries in India connected through the parent network - NICNET.

The project was implemented during 2000 and is accessible over the NIC Intranet at http:JJnews.nic.inJlibraryJnewsnic, and has been running successfully fulfilling the objectives mentioned above.

Advantages

The Web-based news clipping systems have various advantages over traditional clippings services and a few have been listed below:

- instant access of news clippings over the Web through a common user interface;
- global access of news clippings in real time by remote users;
- access to full-text news supplemented with graphics, charts, tables, etc.;
- up-to-the-minute updated access to news;
- dynamic updating of the back-end database from many locations;
- provision of a high level of search options for news archive retrieval;
- instant feedback from users;
- unlimited downloading and printing; and
- environmentally friendly.

System design

To design a Web-based full-text news clipping system we needed various kinds of system components such as hardware, software tools, database tools and Web publishing technologies. As this system has been designed by library professionals who have no formal education or training in computer science we have opted to use such technologies which require less programming skill and were easy to use and implement.

Hardware and software

Information on the hardware and software used is given in Table I.

Database tools

The selection of suitable database tools for library applications depends on the requirements, the kind of applications (bibliographic/numeric/full-text), the application level (data level, data warehouse level), the magnitude of data (number of records, length of fields), the nature of data analysis (basic statistics, i.e. sum, average, mean/advance statistics), comprehensiveness, interoperability, cost, and ease in design and operations (Matoria *et* al., 2002). Keeping in view these points and our requirements to set up a full-text, Web-enabled database at the back-end, we decided to use the Relational Database Model (RDBMS) from among various database tools (pre-relational, relational and object-oriented database models). The reasons for choosing RDBMS included:

- it is the current database model being commonly implemented;
- it provides extremely useful tools for database administration and offers distributed database and distributed processing options;
- it provides referential integrity controls to ensure data consistency;
- it adheres to a powerful query language, i.e. SQL (Structured Query Language) developed by the Microsoft Corporation;
- it is compliant with ODBC/JDBC (Open DataBase Connectivity/Java DataBase

Table I List of hardware and software

S.N.	Item	Quantity	Configuration
01	Web server	01	PIII intel, 8 GBHD, 128 MB RAM, OS = NT4, web server = IIS 4
02	Web clients	04	PIII intel, 4 GBHD, 64 MB RAM, OS = Winn 98
03	Scanner	01	ScanJet 3300 C, 300/600 dpi
04	OCR	01	Fine reader 5.0

Connectivity) and other database interface tools;

- it works in a client/server mode and is thus suitable for the Web environment;
- it is easy to use, develop graphical design interfaces and requires less programming skill.

Our choice of SQL Server

It was observed that the common RDBMS software (MS Access, Oracle, DB2, File Maker Pro, Informix, MySQL, mSQL, Sybase, etc.) tools support up to 255400 characters field length and are thus not suitable for setting up a full-text news clippings database. However, these database tools have a Memo field where a huge amount of data can be stored, but this field is not searchable and so is of limited use for this application. We therefore opted to use Microsoft SQL Server v.7 as the back-end solution for our NewsNIC project as it fulfills our need to design this full-text news clipping system. This database tool has the following unique features:

- It supports searchable field length up to 8,063 characters of data which is enough to store full-text news items (Gunderloy and Chipman, 1999). It has been observed during the last few years of the successful running of the NewsNIC project that 98 per cent of the news items selected are below this size limit. News items that are bigger than this are not stored in the database and instead are stored in a linked Web directory.
- It supports a full-text search facility through the Microsoft search service which offers the ability to issue queries against character data stored in a field. It is based on the full-text index generated and maintained by the SQL server full-text engine. The full-text index so generated can be updated automatically by scheduling the task once with the scheduler, an inbuilt feature of SQL Server.

Database structure

The NewsNIC database at the back-end designed using MS SQL Server 7 has a very simple structure containing two main tables only. The Newspapers table contains details of individual newspapers while another table, News, contains details of the news items to be published. Importantly, the Details field of the News table stores the full-text news in HTML format. Both the tables contain one primary key field to enforce the referential integrity and to make relationships between the tables. Figure 1 shows the structure of the database.

Web authoring tools

The home page for NewsNIC was designed using FrontPage 2000 and Visual InterDev, both software products from the Microsoft Corporation. Both of these are easy to use and provide good administrative plans to maintain the Web site. The site contains two kinds of Web pages, i.e. HTML pages and ASP (Active Server Pages).

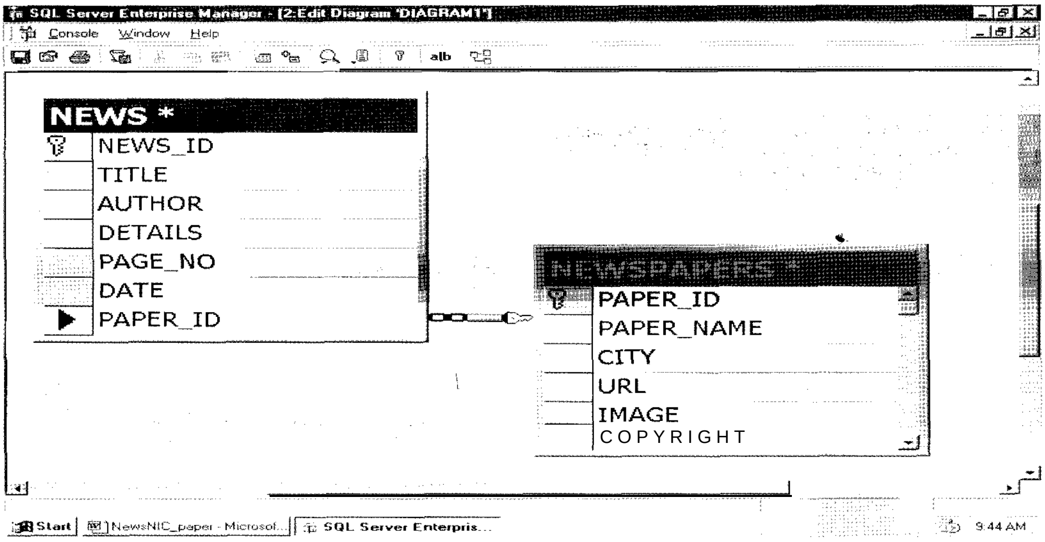
Scripting tools

For publishing and accessing news items direct from the back-end database dynamically, we have used ASP Web server technology developed by the Microsoft Corporation. ASP provides a compile-free application environment where HTML codes as well as scripts (VB Scripts/Java Scripts/ Jscript) are written in the same ASP page. ASP enables server-side scripting for IIS (Internet Information Server from Microsoft Corporation) and thus creates dynamic and powerful Web-based solutions. Moreover, ASP technology is easy to use, requires less programming skill and is therefore suitable for non-programmers.

NewsNIC workflow

The following steps are taken while developing NewsNIC.

Figure 1 Database structure


SQL Server Enterprise Manager - [2:Edit Diagram 'DIAGRAM1']
Console
Window
Help
Start
NewsNIC_paper - Microsof...
SQL Server Enterpris...
9:44 AM

NEWS *
- NEWS_ID
- TITLE
- AUTHOR
- DETAILS
- PAGE_NO
- DATE
- PAPER_ID

NEWSPAPERS *
- PAPER_ID
- PAPER_NAME
- CITY
- URL
- IMAGE
- COPYRIGHT

Selection and marking of news items

The input for the NewsNIC system starts with the selection and marking of useful news items from newspapers. In terms of coverage, we select only IT related news which contains information about our parent organisation, department and ministry as well as the latest developments in India and around the world. In fact, whenever the same news story is published by more than one newspaper we select all the stories as there are always different viewpoints presented by different newspapers. In terms of the number of newspapers we cover under the service, we subscribe to 23 leading national dailies published in various parts of India.

Scanning/OCR/HTML conversion

The selected news items are then scanned and OCRed to convert into HTML format by using an HTML template as seen in Figure 2. The use of this template serves two purposes. First, it is used to maintain uniformity in display, and second the template contains metatags to record the "news elements" such as title, author, date, paper code and pagination. Later, the news elements in the metatags are used to read / write direct to the corresponding fields in the database by the uploading program.

However, nowadays all the newspapers covered in our service are available on the Web, where the news items are already in HTML format. This facilitates the copying and downloading of selected news items from the Web and thus eliminates the tricky and error-prone job of scanning and OCR. We just insert the metatags in the HTML files downloaded from the Web edition and then upload them in the database using an upload program.

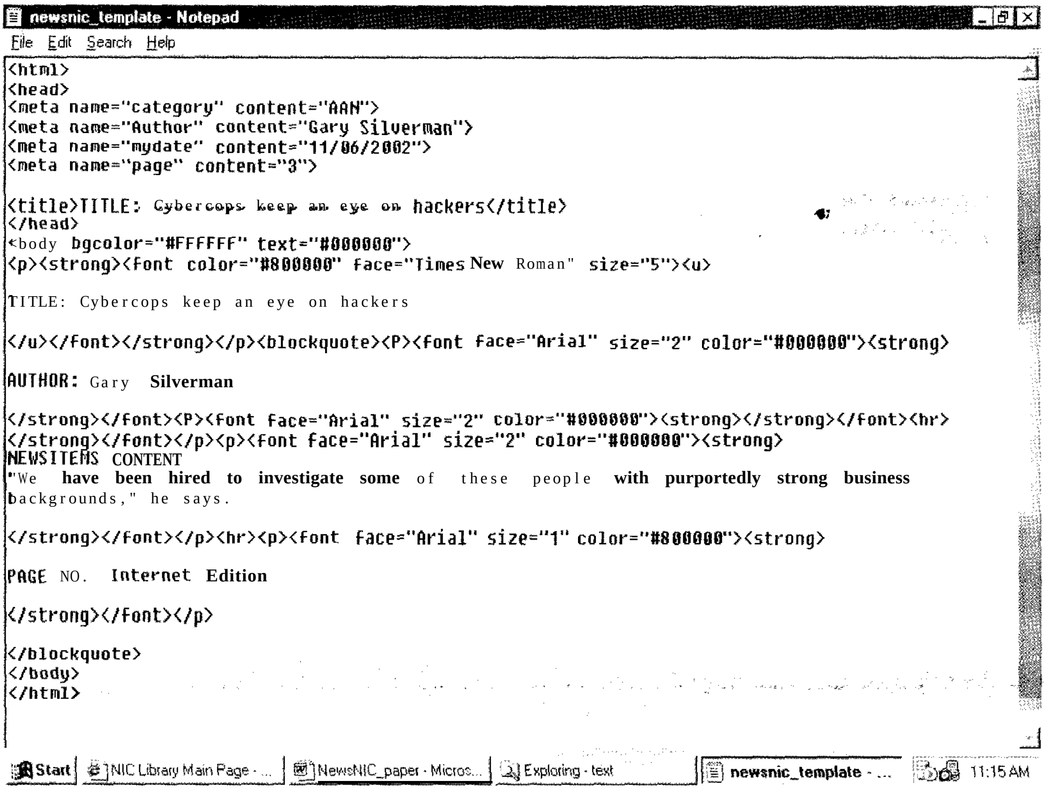
News uploading to database

Updating the database daily with manual methods by keying-in the elements of the news clippings is not possible. Therefore, to simplify the database updating we have designed a News Upload Program as a sub-set of the NewsNIC system. This program reads the news elements, i.e. title, author, date, embedded in the metatags from the HTML files (converted news items in digital form) and writes these elements directly to the corresponding fields in the database.

The News Upload Program is an ASP page containing Virtual Basic codes (FSO, i.e. File Systems Objects) to read, validate and write the news elements from the metatags embedded in the HTML files, and to copy these elements to corresponding fields, thus updating the NewsNIC database at the back-end. Simultaneously, the full-text news items (complete HTML files) are copied into a corresponding field in the database. Later, this field (named as Details) is used for full-text queries as it contains the news clippings in digital form.

XML (Extensible Markup Language), another standard from W3C (Worldwide Web Consortium) as a document description and data exchange format, may be more useful for our project where XML tags describe the structural components (e.g. author, date, etc.) and thus can be used for automatic

Figure 2 HTML **template**


newsnic_template - Notepad
File Edit Search Help
Start
NIC Library Main Page - ...
NewsNIC_paper - Micros...
Exploring - text
newsnic_template - ...
11:15 AM

```
<html>
<head>
<meta name="category" content="AAN">
<meta name="Author" content="Gary Silverman">
<meta name="mydate" content="11/06/2002">
<meta name="page" content="3">

<title>TITLE: Cybercops keep an eye on hackers</title>
</head>
<body bgcolor="#FFFFFF" text="#000000">
<p><strong><font color="#800000" face="Times New Roman" size="5"><u>

TITLE: Cybercops keep an eye on hackers

</u></font></strong></p><blockquote><P><font face="Arial" size="2" color="#000000"><strong>

AUTHOR: Gary Silverman

</strong></font><P><font face="Arial" size="2" color="#000000"><strong></strong></font><hr>
</strong></font></p><p><font face="Arial" size="2" color="#000000"><strong>
NEWSITEMS CONTENT
"We have been hired to investigate some of these people with purportedly strong business
backgrounds," he says.

</strong></font></p><hr><p><font face="Arial" size="1" color="#800000"><strong>

PAGE NO. Internet Edition

</strong></font></p>

</blockquote>
</body>
</html>
```

metadata extraction. In fact, XML provides a universal format for storage and delivery of information and thus eliminates the need for proprietary data formats and the problems associated with converting one type of data to another (Banerjee, 2002).

Automatic indexing of news items

MS SQL Server has an inbuilt full-text engine which, on uploading the news items to database, automatically updates the full-text index as scheduled by the scheduler. In the full-text index so generated, we have included the Details field which stores the news clippings (HTML files) in full. This field contains the news in digital form and is used for full-text searching by matching the word(s) present therein.

The work flow for the production of NewsNIC is shown in diagrammatic form in Figure 3.

Publishing news items on the library Web server

On uploading the news clippings to the database at the back-end, these news clippings become available on the library Web server and thus are accessible on the NewsNIC homepage (as shown in Figures 4 and 5) under various options/links such as latest news, news archives, date-wise search. These links execute the ASP codes embedded within HTML pages which fetch the results from the back-end database. Initially, the results pages display the headlines of the news items along with the paper name and date. The headlines displayed dynamically from the database contain hyperlinks to the corresponding news in full.

Archive search

To search the news archives we have designed an interface for users whereby they can search news by title, author, date, paper name, etc. as can be seen in Figures 6 and 7. In addition, users can search the full-text news by using Boolean operators. The search interface consists of two tiles, i.e. a search form (HTML tile) to submit the user input and a results page (ASP page) which holds the results set. The results display only the headlines along with date and newspaper name of the news items with hyperlinks to the full news.

Figure 3 NewsNIC **work flow**

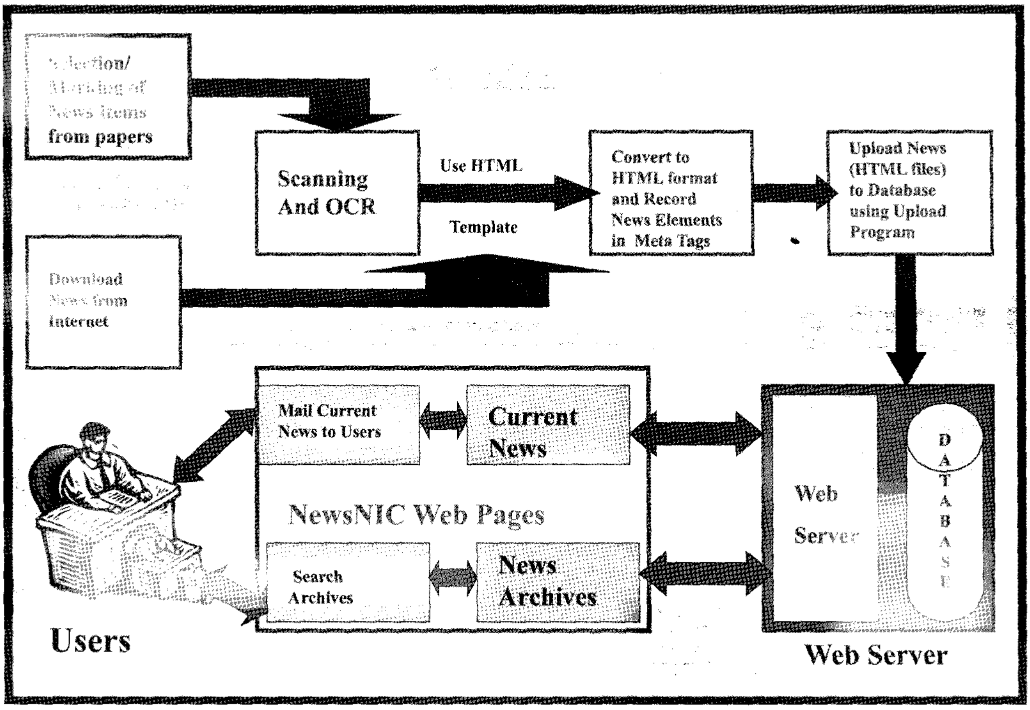
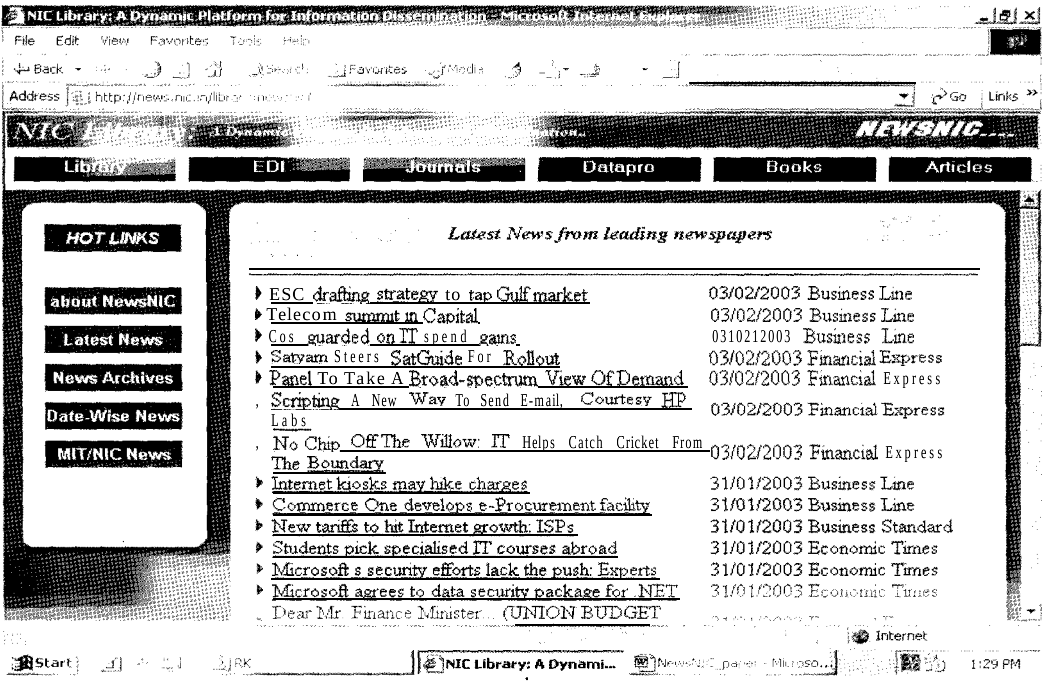


Figure 4 An example of a latest news screen


NIC Library: A Dynamic Platform for Information Dissemination - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://news.nic.in/library/newsnic/
Go Links
NIC Library
NEWSNIC
Library
EDI
Journals
Datapro
Books
Articles
HOT LINKS
about NewsNIC
Latest News
News Archives
Date-Wise News
MIT/NIC News
Internet
Start
RK
NIC Library: A Dynami...
NewsNIC_paper - Microso...
1:29 PM

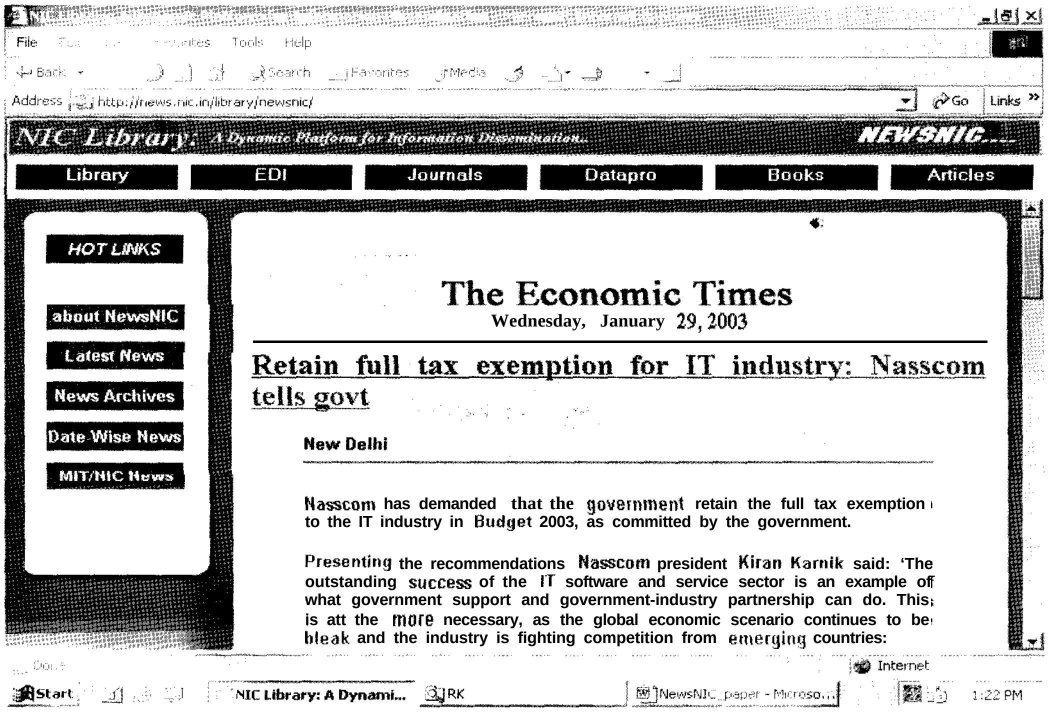
Latest News from leading newspapers

ESC drafting strategy to tap Gulf market	03/02/2003	Business Line
Telecom summit in Capital	03/02/2003	Business Line
Cos guarded on IT spend gains	0310212003	Business Line
Satyam Steers SatGuide For Rollout	03/02/2003	Financial Express
Panel To Take A Broad-spectrum View Of Demand	03/02/2003	Financial Express
Scripting A New Way To Send E-mail, Courtesy HP Labs	03/02/2003	Financial Express
No Chip Off The Willow: IT Helps Catch Cricket From The Boundary	03/02/2003	Financial Express
Internet kiosks may hike charges	31/01/2003	Business Line
Commerce One develops e-Procurement facility	31/01/2003	Business Line
New tariffs to hit Internet growth: ISPs	31/01/2003	Business Standard
Students pick specialised IT courses abroad	31/01/2003	Economic Times
Microsoft s security efforts lack the push: Experts	31/01/2003	Economic Times
Microsoft agrees to data security package for .NET	31/01/2003	Economic Times
Dear Mr. Finance Minister... (UNION BUDGET	[illegible]	[illegible]

Delivery of news by e-mail

On updating the NewsNIC database daily up to 12 o'clock, the headlines of current news with hyperlinks to full news are then e-mailed to library members and they can click the individual headline to display the full news through an Internet browser. The e-mail addresses of the members are stored in a database. To send the news headlines direct from the database, we have again used ASP technology where CDONTS (Collaboration Data Objects for Windows NT Server) of Visual Basic has been used. This object works on the SMTP (Simple Mail Transfer Protocol) service of the Internet Information Server and helps in sending the e-mail in bulk to many users with a single click, and users receive and read the mail as shown in Figure 8.

Figure 5 An example of some full-text news



The Economic Times
Wednesday, January 29, 2003

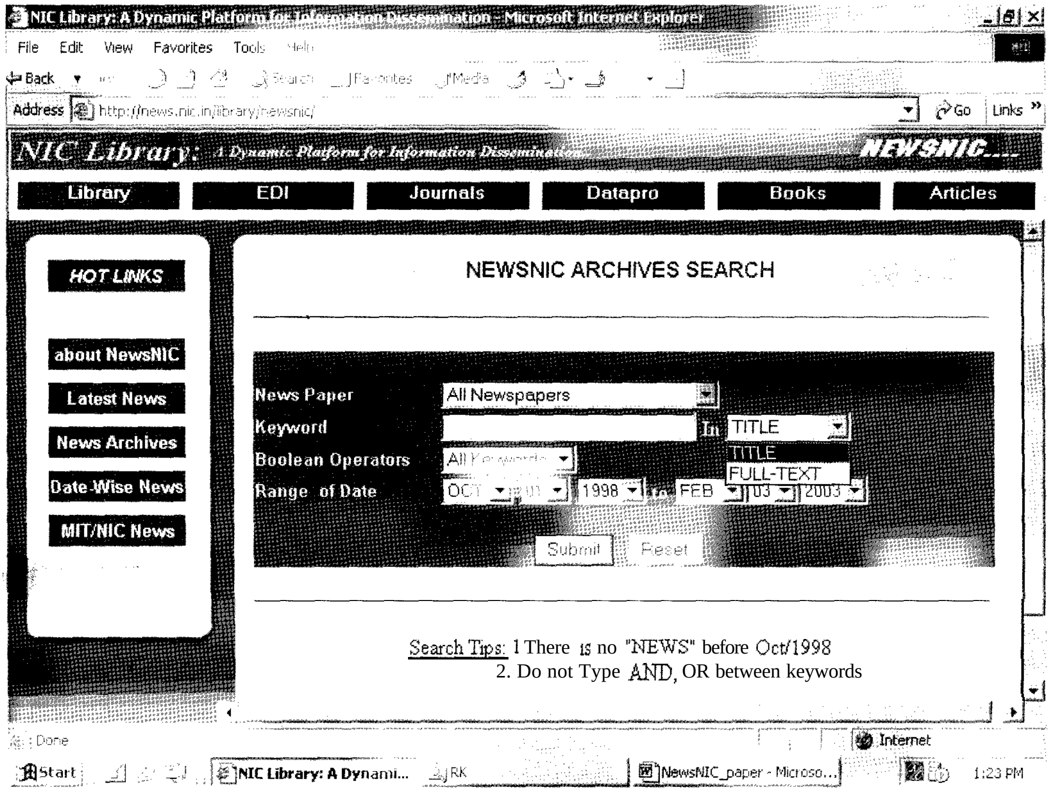
Retain full tax exemption for IT industry: Nasscom tells govt

New Delhi

Nasscom has demanded that the government retain the full tax exemption to the IT industry in Budget 2003, as committed by the government.

Presenting the recommendations Nasscom president Kiran Karnik said: 'The outstanding success of the IT software and service sector is an example of what government support and government-industry partnership can do. This is att the more necessary, as the global economic scenario continues to be bleak and the industry is fighting competition from emerging countries:

Figure 6 Archives search form



NEWSNIC ARCHIVES SEARCH

News Paper	All Newspapers		
Keyword		In	TITLE TITLE FULL-TEXT
Boolean Operators	All Keywords		
Range of Date	OCT 01 1998	to	FEB 03 2003

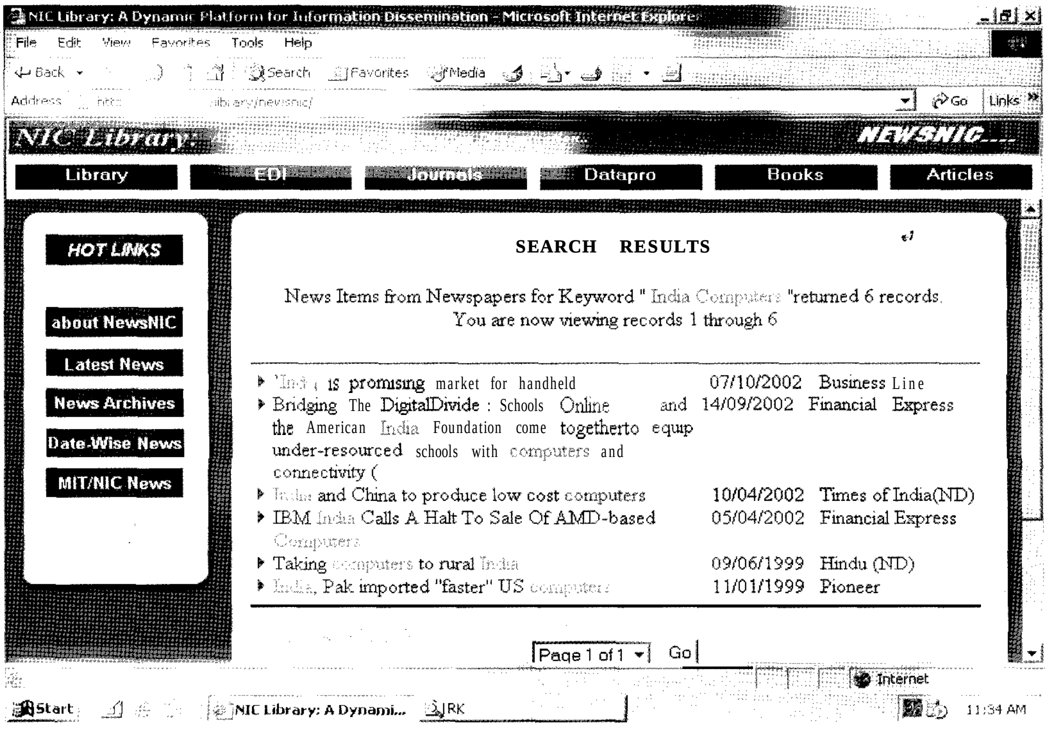
Submit Reset

Search Tips: 1 There is no "NEWS" before Oct/1998
2. Do not Type AND, OR between keywords

Copyright issues

In a general sense, the *Indian Copyright Act* (1957, amendments – 1999) provides various provisions to protect the rights of copyright holders and restricts the commercial use of an original work by any other agencies. However, it advocates the fair use of reading materials by libraries and information centres for educational and research purposes. In view of

Figure 7 Results from an archives search


NIC Library: A Dynamic Platform for Information Dissemination - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http: library/newsnic/ Go Links
NIC Library:
NEWSNIC
Library EDI Journals Datapro Books Articles
HOT LINKS
about NewsNIC
Latest News
News Archives
Date-Wise News
MIT/NIC News
Internet
Start NIC Library: A Dynami... RK 11:34 AM

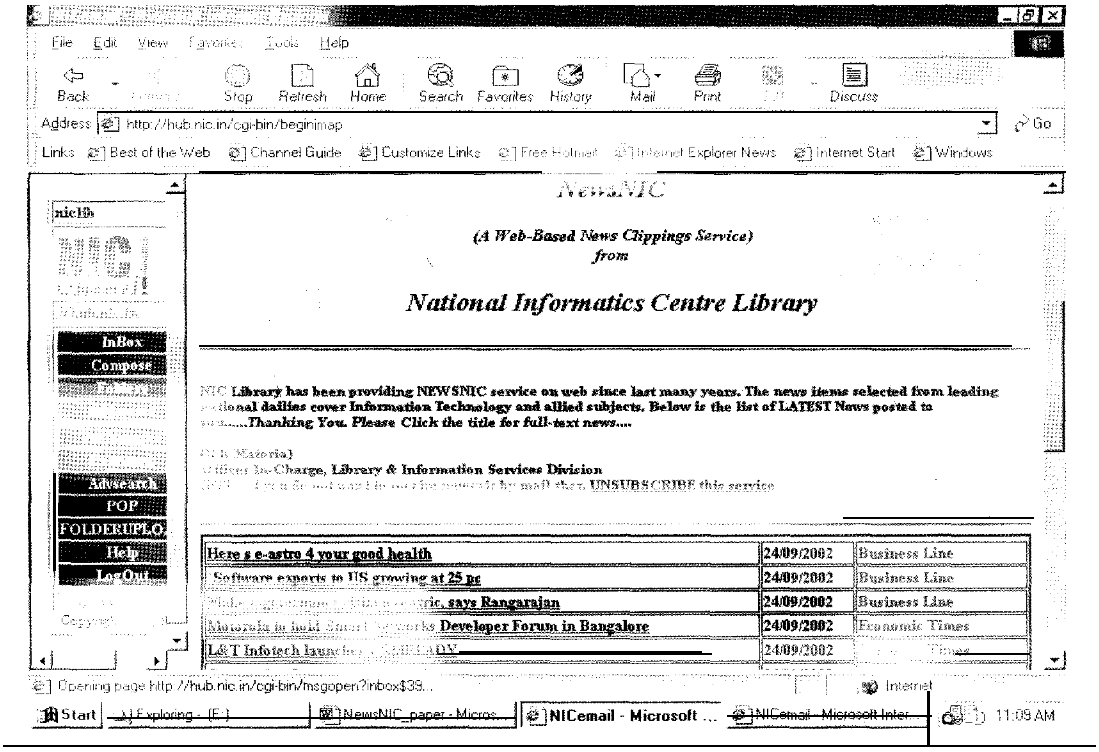
SEARCH RESULTS

News Items from Newspapers for Keyword " India Computers "returned 6 records.
You are now viewing records 1 through 6

▸ India is promising market for handheld	07/10/2002	Business Line
▸ Bridging The DigitalDivide : Schools Online and the American India Foundation come togetherto equp under-resourced schools with computers and connectivity (	14/09/2002	Financial Express
▸ India and China to produce low cost computers	10/04/2002	Times of India(ND)
▸ IBM India Calls A Halt To Sale Of AMD-based Computers	05/04/2002	Financial Express
▸ Taking computers to rural India	09/06/1999	Hindu (ND)
▸ India, Pak imported "faster" US computers	11/01/1999	Pioneer

Page 1 of 1 Go

Figure 8 Delivery of latest news by e-mail


File Edit View Favorites Tools Help
Back Stop Refresh Home Search Favorites History Mail Print Discuss
Links Best of the Web Channel Guide Customize Links Free Hotmail Internet Explorer News Internet Start Windows
niclib
InBox
Compose
Advsearch
POP
FOLDERUPLO
Help
LogOut
Internet
Start Exploring - (E:) NewsNIC_paper - Micros... NICemail - Microsoft ... NICemail - Microsoft Inter... 11:09 AM

Address http://hub.nic.in/cgi-bin/beginimap

NewsNIC

(A Web-Based News Clippings Service)
from

National Informatics Centre Library

NIC Library has been providing NEWSNIC service on web since last many years. The news items selected from leading national dailies cover Information Technology and allied subjects. Below is the list of LATEST News posted to you.....*Thanking You. Please Click the title for full-text news....*

(R K Matoria)
Officer In-Charge, **Library & Information Services Division**
If you do not want to receive news by mail then UNSUBSCRIBE this service

Here s e-astro 4 your good health	24/09/2002	Business Line
Software exports to US growing at 25 pc	24/09/2002	Business Line
[illegible] ric, says Rangarajan	24/09/2002	Business Line
Motorola to hold Smart Networks Developer Forum in Bangalore	24/09/2002	Economic Times
L&T Infotech launches [illegible]	24/09/2002	[illegible] Times

Opening page http://hub.nic.in/cgi-bin/msgopen?inbox$39...

the right of information and disseminating the information, the concept of fair use has been conceived and legalised by the international communities.

In fact, the newspapers which we include in our service have already set up their own Web sites with free access for the general public over the Internet. The NewsNIC contents available through the library Web server are for internal use by NIC officials only over the NIC Intranet. We have included the copyright statement on each of the news items selected with their corresponding copyright holders.

Conclusion

The usefulness of information in a library depends on how quickly and easily users get the information. IT helps in achieving this purpose and information can be acquired, processed, and distributed quickly, economically and efficiently. Moreover, with the use of Web and database technology, it has become possible now to deliver the information dynamically and instantly to remote users over the Web.

Keeping in view the above facts, the National Informatics Centre (NIC) Library has implemented the Web-based, full-text news clipping system over its Intranet. The system is running well, being appreciated and used extensively by library users all over the country. This system helps in providing the latest news and archives to users posted at remote corners of the country, making them feel connected to the rest of India and the world by reading the news about IT and allied subjects. Moreover, the success of this project sets an example of excellence whereby library professionals have taken the initiative to design and implement such an advanced system.

The authors hope that this case study will be useful in the design and implementation of a Web-based news-clipping service in other libraries.

References

Banerjee, K. (2002), "How does XML help libraries?", *Computers* in Libraries, Vol. 22, No. 5, available at: www.infotoday.com/cilmag/sep02/Banerjee.htm

ePRO System Limited (2002), "Web-based news clipping library imaging system", developed by ePRO Systems (HK) Limited, available at: www.epro.com.hk/News%20Clipping%20Library%20Info%20System.pdf

Gunderloy, M. and Chipman, M. (1999), *SQL Server 7 in* Record Time, BPB Publications, New Delhi.

Matoria, R.K. and Upadhyay, P.K. (2002), "Design and development of Web-enabled databases in libraries with special reference to RDBMS: selection of tools and technologies", DESIDOC Bulletin of Information Technology, Vol. 22 No. 4, pp. 9-16.

Mühlberger, G. (1999), "Digitization of newspapers clippings: the LAURIN project", *RLGDigi* News, Vol. 3 No. 6, pp. l-20, available at: www.rlg.org/preserv/diginews/diginews3-6.html#feature

NICNET (2002), "A profile", available at: www.home.nic.in/htm/nicnet.htm

Pownikar, S., Pusalkar, S., Deepa, V. and Karkhanis, S.N. (2002), "E-clippings: the electronic newspaper clipping service of C-DAC Technical Library -- a case study and lessons learned", in Parthan, S. and Jeevan, V.K.J. (Eds), Information Management in e-Libraries, Proceedings *of the* National Conference */-/e/d in IIT* Khargpur, India, *Feb.* 26-27, 2002, Allied Publishers, New Delhi, pp. 37-42.

Singh, M., Singh, S.K., Gupta, R. and Kumar, V. (2000), "E-clippings: paperless nuclear science and technology news retrieval and archival application", in Kaul, H.K. (Ed.), *NACLIN 2000:* Library and Information *Networking,* Proceedings of the DELNEJ *Conference held in Chennai, India, December 24-25* 2000, Developing Library Network, New Delhi, pp. 129-37.