DESIDOC Bulletin of Information Technology , Vol. 25, No.3, May 2005, pp. 9-17

Application of Greenstone Digital Library (GSDL) Software in Newspapers Clippings

Sharad Kumar Sonkar, Veena Makhija, Ashok Kumar, Dr Mohinder Singh

Abstract

With the 'digital library' emerging as a trend in the present era, the information society of today suffers not from the lack but from a surplus of information. Distinguishing relevant from non-relevant information is therefore, one of the main tasks of librarians. The development of information technology in the field of digital libraries has attracted many research efforts during the last years. Many interesting projects have been started, dealing with the various open issues arising in this field addressing the challenges such as metadata selection, preservation, technology obsolescence, and copyright issues, etc.

DESIDOC has taken several initiatives in this direction. Building a 'digital library of newspaper clippings' is one such effort in this direction. This articles discusses in detail all the issues related to the development of digital library of newspaper clippings and implementation of 'Greenstone Digital Library' software in developing such collection.

1. INTRODUCTION

In this so called 'age of information' we live in, individual people, government agencies, and private corporations have an almost unlimited need for news and information about the topics that affect their daily lives and transactions. Information is truly power, and especially for research and businesses. The more information they have, the better off they are.

For more than 100 years an enormous number of libraries, archives, companies, NGOs and governmental institutions have been collecting newspaper clippings (NPC) and maintaining clipping archives. With the emergence of online-editions of newspapers and networks of such electronic editions some of them have closed their paper-based collections. But most of the other institutions still maintain their paper-based clipping collection, and they will continue to do so in the future. The reasons for this are: (a) their users are not only interested in the pure text of a newspaper article but demand a higher level of authenticity; and (b) the typography, layout, pictures that come with it, and the impression of the whole page attracts them.

As long as print editions are the main distribution media of newspapers, it will make sense to collect clippings of the print edition, because they have this value added to them—either by managing them out manually, or by scanning them and cutting them electronically. But this is not the only reason why we are optimistic about the future of clipping archives.

Another reason lies in the general role research libraries will play in the near future: The information society of the future will not suffer from a lack but from a surplus of information. Distinguishing relevant from non-relevant information will, therefore, be

one of the main tasks of librarians in the future. Since the very beginning, it has been the job of archivists to select from a wide range of newspapers, the specific articles related to the subject of the collection. The development of information technology in the field of digital libraries has been attracting many research efforts during the last years. Many interesting projects have been started, dealing with the various open issues arising in the field. In India NCSI has setup the open access archive for Indian Institute of Science research publications and DRTC has setup an open access digital library for library and information science literature published in India. However, no project has specifically taken into account the problem of building a digital library of NPC.

It is well known that a huge part of cultural knowledge is stored in the newspapers of yesterday. Since newspapers are not always easily accessible, special clipping archives were created in the 20th century. Researchers are interested in newspaper information from these archives because of the librarian's efforts of selecting, cutting and indexing the articles. In order to maintain their important position in the information market, clipping archives should be able to integrate their special skills (such as professional knowledge and experience in gathering and treating newspaper information) into the new technologies of the information society. The majority of users, who may be able to use a database but who are no specialists, will prefer a system where library professionals have reduced complexity to a degree which allows non-professional users to access relevant information in a comfortable way.

2. NEWSPAPER CLIPPING PROCESS

To continue the tradition of collecting and providing access to newspaper clippings in the networked environment, the DESIDOC embarked upon the digitisation of NPC project in 2000. A model for access and retrieval of digitized NPC was developed, it can be accessed through DRONA (DRDO Rapid Online Network Access). DRONA provides simple and advance searching. DRDO researcher can search the information in various field like Title, Category, Name of Newspapers, Date of Publication. We are providing NPC service both in paper and electronic form. It includes seventeen national newspapers in Hindi and English.

Figure 1 shows the process of providing access to NPC in networked environment.

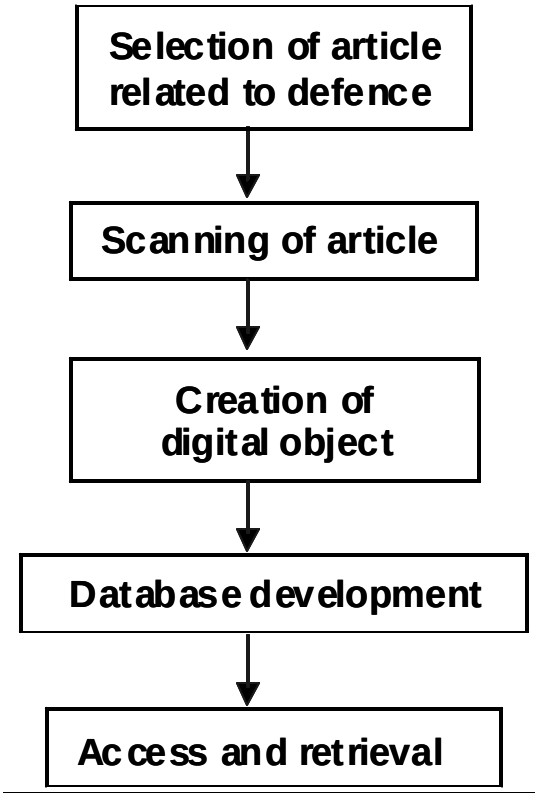

Figure 1. News clippings workflow

The various steps of this process are discussed below:

(a) Selection

Scientists scan the content of newspapers and select only those article pertaining to defence technologies, weapons, defence policies, DRDO and International relations. They go through all the newspapers and mark relevant articles every day. (Figure 2)

(b) Scanning

Selected articles are scanned and saved in a required format Indicating the name of the newspaper and publication date.

Figure 2. Scanning of news clippings

(c) Creation of Digital Object

Scanned articles are used as digital object, we create digital object in format like jpg, bmp, or doc. In order to create a doc file, we perform the OCR (optical character recognition) processing and OCR correction using Abbey Fine Reader.

(d) Database Development

These procedures are mainly done in the background. It requires bibliographic information (title of the article, name of news papers, date of the newspapers, category, file name) recorded at this point. (Figure 3)

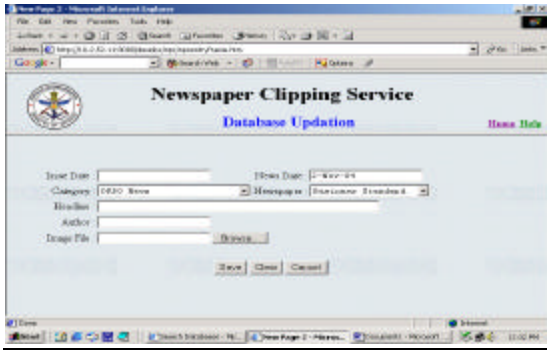

Figure 3. Database development

(e) Access and retrieval

The NPC service provides browsing and search facility in various fields. Advance searching provides search facility in various fields like headline and author, etc. Users can also precise their search results by limiting the time period, etc. (Figure 4)

However, we encountered one limitation during the search of news items the resource locators were provided to users in the form of keywords in title, name of author, categories, newspaper name and the publication date, but it didnot provide sufficient input to the users to retrieve relevant data, specially during retrospective search of news items. To overcome this barrier we decided to go for full-text searching of NPC database.

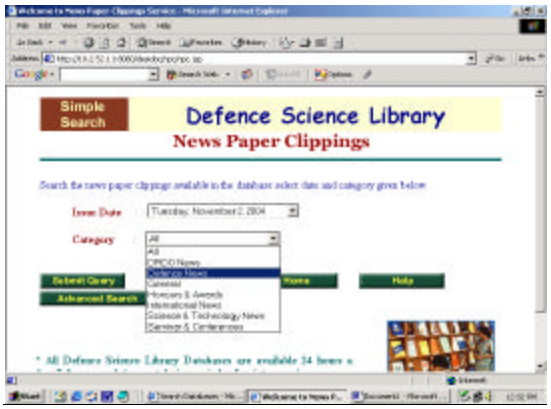

Figure 4. Access and retrieval

3. GREENSTONE DIGITAL LIBRARY SOFTWARE

In order to develop a digital library of NPC at DESIDOC, which allows full text searching of data we thought of opting for Greenstone digital library software, which is an open-source system for the construction and presentation of information collections. It builds collections with effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. Moreover, they are easy to maintain. Collections of information comprise large numbers of documents (typically several thousand to several million), and a uniform interface is provided to them. It is used internationally. Greenstone supports collections in many different languages. Greenstone CD-ROMs have been published by the United Nations and other humanitarian agencies for distribution in developing countries.

According to Greenstone website (www.greenstone.org), “The aim of the software is to empower users, particularly in universities, libraries, and other public service institutions, to build their own digital libraries. Digital libraries are radically reforming how

information is disseminated and acquired in UNESCO's partner communities and institutions in the fields of education, science and culture around the world, and particularly in developing countries. We hope that this software will encourage the effective deployment of digital libraries to share information and place it in the public domain."

4. SPECIAL FEATURES

The GSDL software is an open source software available from the New Zealand Digital Library (www.nzdl.org) under the terms of GNU, general public library license. Greenstone CD-ROMs have been published by the United Nations and other agencies for distribution in developing countries. Some of the features of the software are:

- It suits both Windows (3.1/3.11, 95/98/me, NT/2000) and Unix (Linux Sun OS) any of these systems can be used as a web server.
- The administration function build in it enables the items to authorise new users to build collection, protect documents so that they can only be accessed by registered users on presentation of password.
- It build collection with effective full-text searching and metadata-based browsing facilities. Collection containing millions of documents, up to several gigabytes can be built. Full-text searching is fast because compression is used to reduce the size of the indexes and text users can browse the list of authors, titles, date, class no., etc.
- Plug Ins can be written to accommodate new document types, the collection can contain pictures, music, audio, video clips, etc. It also supports multilingual documents.
- Collection can be updated and new one brought online any time with out bringing down the system.

The Z39.50 protocol is supported for accessing external servers and for presenting Greenstone collection to external clients.

5. SOFTWARE REQUIREMENTS

- OS Windows/Linux
- Apache web server/IIS
- PERL
- Java 2 Runtime Environment 'version 1.4.2_03'
- Web browsers—Netscape Navigator or Internet Explorer
- GSDL 2.41

6. STEPS FOR INSTALLATION OF SOFTWARE

6.1 Installation of GSDL on Windows [web library]

The following steps are needed for installation:

1. Install the web server IIS/Apache
2. Install the Java 2 Runtime Environment from the internet (latest version)
3. After installing J2RE, Go to GSDL Folder "gsdl-2.41-win32" (Setup file) from the internet

   "MyComputer-GSDL-"gsdl-2.41-win32. exe"
4. Choose Setup Language. English [United States] is the default
5. The InstallShield Wizard will begin the installation of GSDL software. Click<next>
6. Accept all the terms of license agreement by clicking on <yes> button
7. Choose the type of installation you need and choose the collection/s that you want to be installed.
8. Set the admin password

(The above step will install web library edition of GSDL and any other sample collection/s and/or GSDL documentation, CD exporting function depending on what was checked or unchecked) (Figure 5)

9. Check the Greenstone Directory Structure:

   D: gsdl/Collect

   Cgi-bin

   Micros

   Gli etc.

We should create virtual directories for GSDL home and CGI executable directory:

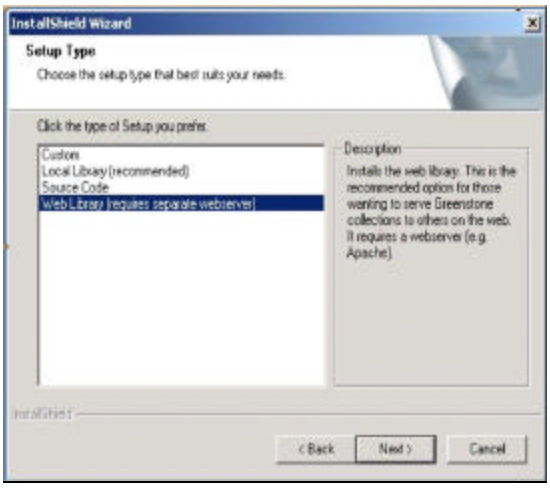

Figure 5. Installation wizard

6.2 Creating Virtual Directory for GSDL Home

Start-Programs-Administrative tools-Internet Services Manager-Click on Hostname/Computer Name-Right Click on Default Web Site-Select New-Click on Virtual Directories-Click Next Button-Provide Virtual Directory Name (gsdl) (Figure 6)

- Click on Next Button-Browse and Select GSDL Web Site
- Content Directory (D:\program files\gsdl)

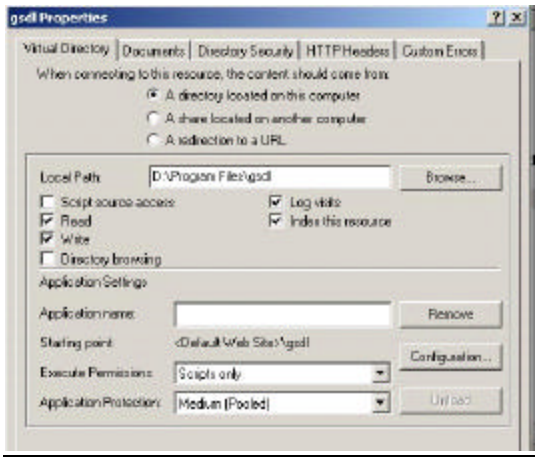

Figure 6. GSDL homepage

6.3 Creating Virtual Directory for CGI Executable Directory

Start-Programs-Administrative tools-Internet Services Manager-Click on Hostname/Computer Name-Right Click on Default Web Site-Select New-Click on Virtual Directories (Figure 7)

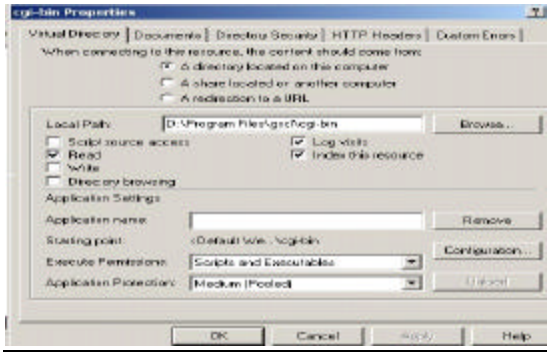

Figure 7. Virtual directory for CGI executable directory

- Click Next Button-Provide Virtual Directory Name (cgi-bin)
- Click on Next Button-Browse and Select GSDL Web Site
- Content Directory (D:\program files\gsdl\cgi-bin)

The above steps complete the installation of GSDL Web Library and also configuration of IIS to be able to recognise GSDL installation directory. The document properties are shown in figure 8.

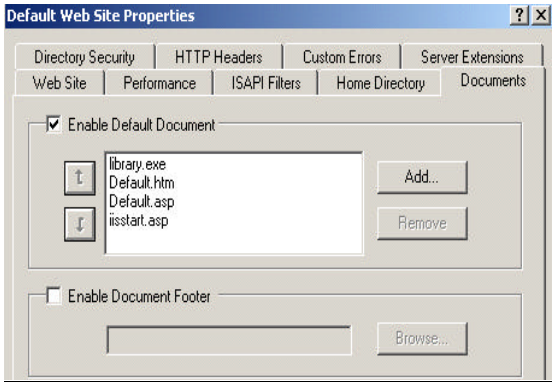

Figure 8. Document properties

6.4 Accessing the Web Library Edition

We can access the web library edition of GSDL from a web browser by using one of the following URLs:

http://localhost/gsdl/cgi-bin/library.exe (or)
http://XXX.XX.XX.XXX/gsdl/cgi-bin/library.exe

7. GREENSTONE LIBRARIAN INTERFACE

The Greenstone Librarian Interface is a tool for collecting and marking up documents, then building digital library collections. It provides access to the GSDL software's functionality from a graphical point. Following steps are involved in creation of digital collection:

7.1 Creation of New Collection

To create a new collection open the 'file' menu and choose 'New' several field need to fill out for your collection, then the following screen will appear. (Figure 9)

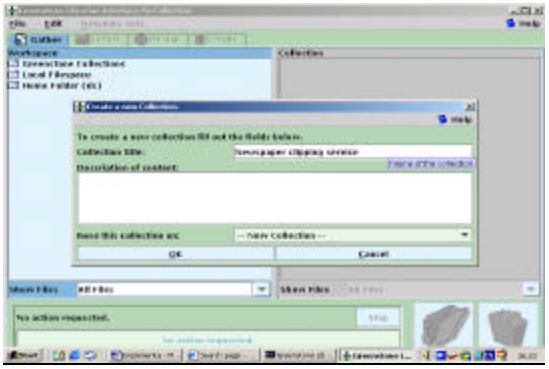

Figure 9. Creation of new collection

7.2 Selection of Metadata

Software provided different type of metadata. In the screen shown in figure 10, we can select the International standard metadata (Dublin core 1.1 (DC)) for the collection.

(a) Gathering

Gather area to select what files to include in the collection you are building. Files can be copied into the collection by dragging and dropping. (Figure 11)

(b) Enrich

Having gathered several files into the collection, now enrich them with additional information called 'metadata'. This section explains how metadata is created, edited, assigned and retrieved, and how to use external metadata sources. (Figure 12)

(c) Design

Once the files are marked up with metadata, you next decide how it should appear to users as a Greenstone collection. What kind of information is searchable? What ways are provided to browse through the

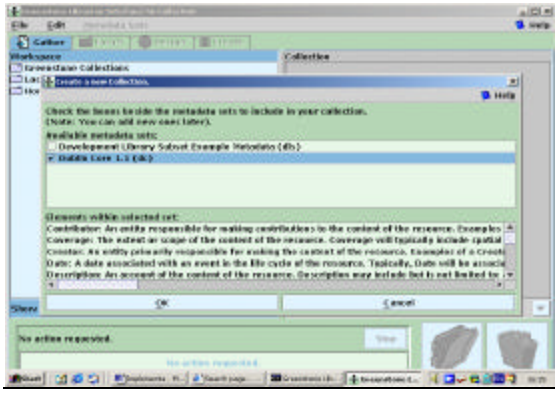

Figure 10. Selection of metadata

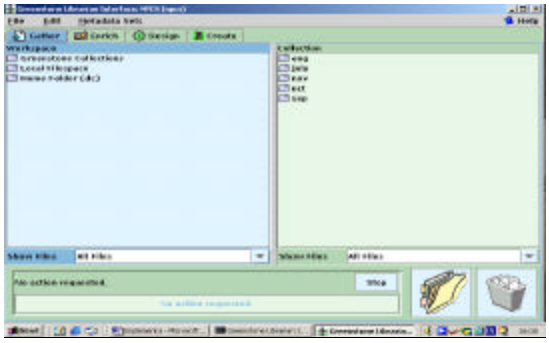

Figure 11. Gathering area

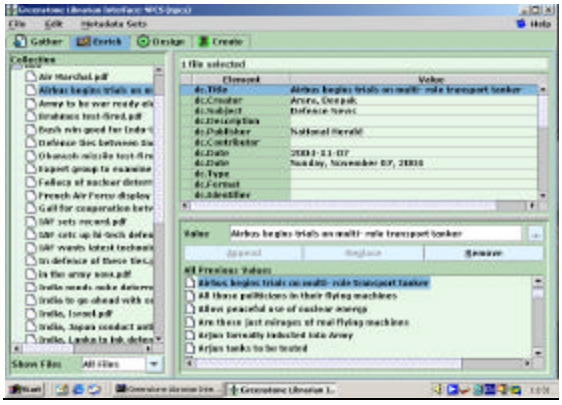

Figure 12. Enriching

documents? What languages are supported? Where do the buttons appear on the page? These things can be customized.

General: Here you can modify the value of your collection which you had filled out at time of creation of new collection.

Search Type: Here you can add search type. If you want to add advance search please enable advance search option.

Search Index: Here you can add search the names of the fields on which you want to build the indexes.

Browsing Classifier: Indexes specify what part of the collection are searchable this section help to add and remove indexes, and set a default index. (Figure 13

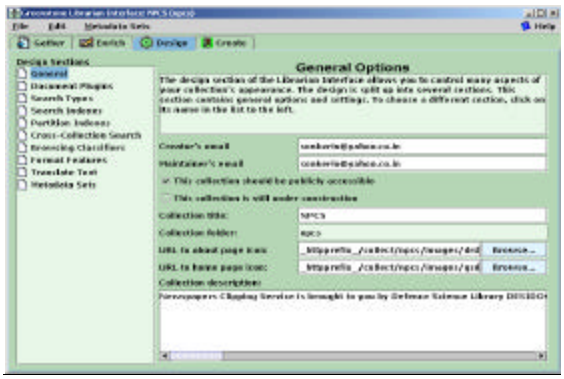

Figure 13. Designing

(d) Create

The first two entries in the list on the left are 'Import' and 'Build', which give settings that apply to the import and build scripts respectively.

8. DIGITAL COLLECTIONS OF NEWSPAPERS CLIPPINGS

- Total 600 newspapers clippings were included.
- All the newspapers clippings are converted into PDF format for more portability.
- Newspapers clippings are organised broadly by the subject category, using 'Greenstone Librarian interface' accompanying software available with GSDL.
- Newspapers clippings are browsable by author, title, subject category, and newspaper names.
- Full-text searching is possible with the following parameters (exact phrase, words in the text). Articles are searchable with simple and advance search interface.
- The search interface can be customized according to the requirements of the users.
- The collection is published in a self installable CD-ROM which can run on any system with minimum configuration.
- At present, size of the collection is 65 MB.
- Collection can be upgraded very easily.

9. SCREEN SHOTS OF THE COLLECTION

Some of the screen shots of the collection.

- Browse by Title (Figure 14)

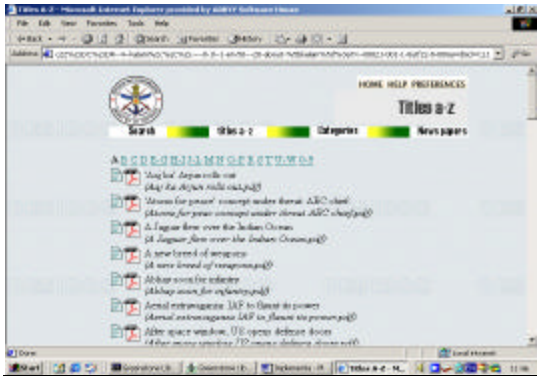

Figure 14. Browse by title

- Browse by Category (Figure 15)

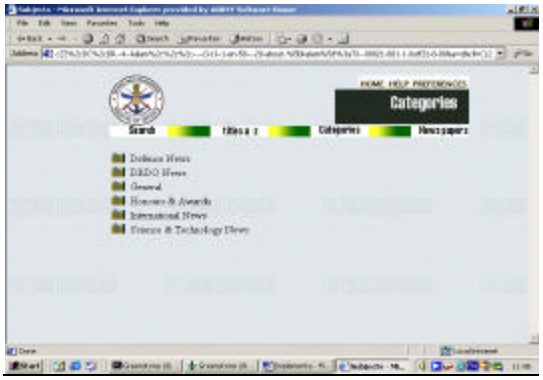

Figure 15. Browse by category

❍ Browse by News papers (Figure 16)

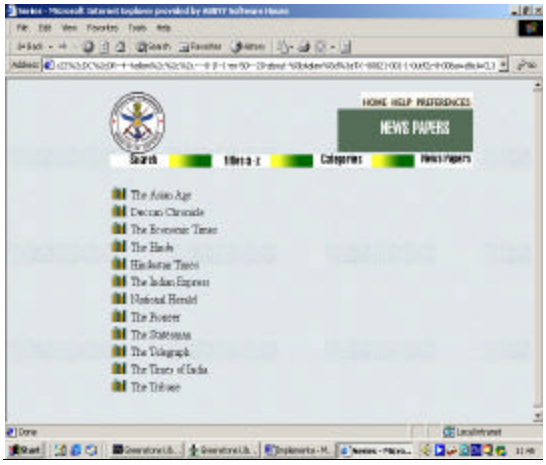

Figure 16. Browse by newspapers

❍ Search facility (Figure 17)

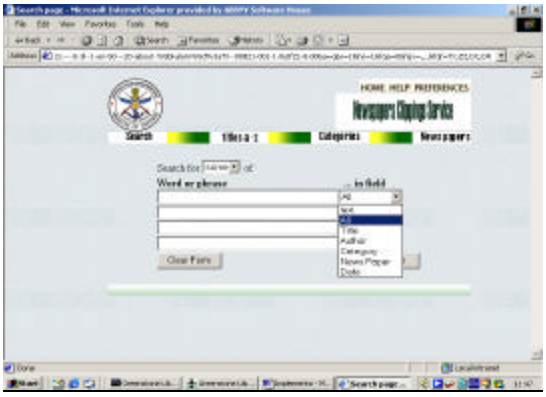

Figure 17. Search facility

❍ Result of retrieved query (Figure 18)

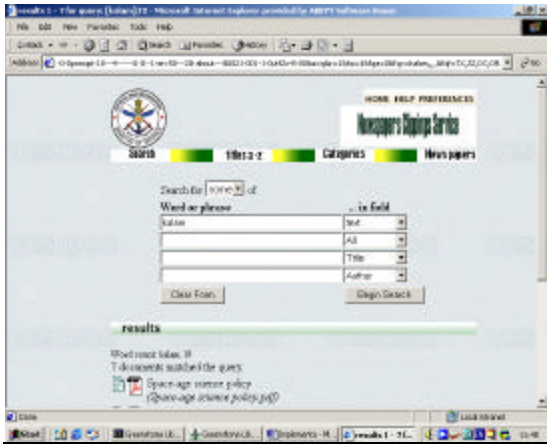

Figure 18. Result of retrieved query

10. OPPORTUNITIES

The web-based NPC systems have various advantages over traditional NPC service and a few have been listed below:

- Providing full-text searching for news archive retrieval
- Global access of news clipping in real time by remote users
- Browsing facility through various field
- Instant access of clipping over the web through a common user interface.
- Up-to-the minute updated access to news
- Dynamic updating of the digitized collection from any location
- Unlimited downloading and printing, and environment friendly.

CONCLUSION

Digital libraries are evolving from an experimentation stage to institutionalisation. Next generation digital libraries will supply a comprehensive range of services on network, not as an institutional prerogative, but as a universal search library. The library that meets the information needs with the collections and services which were previously maintained in-house. The approach towards setting up the digital library standards, software technologies and best practices is definitely taking a shape.

The information professional has to keep constant watch for new developments and noticeable changes in the field of their concern. To cope up with the information needs with speed and relative accuracy and reliability, the digital library has emerged as the most important and reliable resort, and so knowledge discovery in these type of libraries becomes a predominant factor. Greenstone offers scores of collection and represents the cutting edge of digital library research using greenstone as a vehicle for dissemination. It help to create different types of collection like audio. Video, image, text and multimedia collection. The collection of NPC involves a huge ancillary effort in digitising thousand of newspapers clippings.

The digital libraries arena offers a unique challenge to an emerging breed of digital

librarians, to combine principles, practices and tools of information management to create new information product and service. The GSDL offers a dynamic platform for knowledge organization and retrieval.

REFERENCES

1. ePRO System Limited. Web-based news clipping library imaging system. Developed by ePRO systems (HK) Limited, 2000. http://www.epro.com.hk/News%20Clipping%20Library %20Info%20System.pdf (February 13, 2004)
2. Muhlberger, G. Digitization of newspapers clippings: The LAURIN Project. *RLG Digi News,* 1999, **3**(6), 1-20. http://www.rlg.org/preserv/diginews/diginews3-6.html (February 13, 2004)
3 Meitetsu, Haruyama. New services and future challenges of the detached library in the diet: Guidance on library services for "Diet People." *National Diet Library Newsletter*, October 2003, **133**. http://www.ndl.go.jp/en/publication/ndl_newsletter/133/331.html
4. Mühlberger, Günter. Newspaper clippings in a digital world: The LAURIN project. http://www.exploit-lib.org/issue2/laurin/
5. How to start your own home newspaper clipping service (February 13, 2004) http://www.work-at-home-index.net/article335.htm
6. Calvanese, D.; Catarci, T.; & Santucci, G. Building a digital library of newspaper clippings: The LAURIN Project. *In* Proc. of the IEEE Forum on Research and Technology Advances in Digital Libraries (ADL 2000), 2000, pp. 15–26.
7. Maruthu Pandian, B.; Sonker, Sharad Kumar; & Moorthy, R. Creating digital libraries: An experiment with Green Stone Digital Library Open Source Software. *In* Proceedings National Conference on Change Management in Library and Information Centers, *edited by* Ally Sornam, S. & Geetha, V. Tiruchirapalli, Tamil Nadu. 2003.

Contributors: **Sh. Sharad Kumar, Sonker**, STA 'A', Defence Scientific Informatgion and Documentation Centre (DESIDOC), Metcalfe House, Delhi - 110 054.

Smt. Veena Makhija, Sc 'D', DESIDOC, Metcalfe House, Delhi - 110 054.

Sh. Ashok Kumar, Sc 'E', DESIDOC, Metcalfe House, Delhi - 110 054.

Dr Mohinder Singh, Director, DESIDOC, Metcalfe House, Delhi - 110 054.