Bangladesh Journal of Library and Information Science

VUNA: A Prototype for Organisation and Retrieval of Newspaper Items

Goutam Maity[1]

Abstract

This is the age of information explosion. A massive quantity of information is being generated, a large portion of which is found in newspapers. It is posing a unique challenge for information managers for processing information and providing required information to information seekers. This article discusses a newspaper information retrieval system named Vidyasagar University News Archive (VUNA), which has been developed by information specialists for effective management and retrieval of information from newspapers. This article highlights the salient technical issues related to the development of a prototype online digital newspaper information retrieval system suitable for India. The basic architecture, operational features, etc. have been discussed and a comparative picture of respective performance and failures of existing online newspaper archives in India and VUNA also has been drawn.

Keywords: Newspaper archives, Prototype design, VUNA

A large number of newspapers coming regularly with sheer volume and wide variety of information, presently, create information explosion/overload. In consequence, the information workers have been facing a lot of hurdles regarding the storage and retrieval of such information. The increasing need and demand for retrospective use of newspaper information, from among the people involved both with the 'research and development activities' and 'application activities', have compelled them to become more concern towards this issue. Till date, several printed and manual information retrieval systems have come out as solution to this problem. But these systems are gradually becoming out of date, since they offer limited scope and facility for searching and retrieval of information. Hence, the specialist users, now, very little depend on them for having their required information. However, a few online digital archiving initiatives for newspaper items, mostly offered by the newspaper establishments, have been found in India. No doubt, these online systems are far better(at least in terms of searching efficiency), in meeting users' demands, than their printed and manual

[1] *Reader, Department of Library and Information Science, Jadavpur University, Kolkata-700032, India. E-mail-gm_vu@yahoo.co.in*

counterparts. Still these are functioning with many lacunas and thereby have failed to cater for users' demands. Moreover, these are not up to the mark in view of the state-of-art ICT and the latest trends and developments in information storage and retrieval procedure.

The situation, therefore, has created a strong demand upon the information specialists towards designing an efficient and effective newspaper information retrieval system. Keeping this in view, an attempt has been made, here, to design and develop a prototype online digital newspaper information retrieval system, suitable for India.

Objective

This prototype news archive is named as Vidyasagar University News Archive (VUNA). Apart from supporting all the facilities to overcome limitations of web based digital newspaper archives in India, it extends measures to achieve following features.

1) Should be amenable for multiple languages;
2) Should be cost effective and time efficient to the users;
3) Should be simple and easy to search but not at the cost of hampering retrieval of relevant items;
4) Should design architecture, completely based on open source software; and thereby it should provide complete freedom in customization and source code modification;
5) Should have web integrated architecture allowing seamless access over Internet, intranet and extranet;
6) Should support the international metadata harvesting protocol - OAI/PMH;
7) Should support general metadata schema like Dublin Core and domain specific metadata schema like NewsML;
8) Should use a relational database (Managing Gigabyte) at the backend for efficient data management;
9) Should provide excellent control over indexing of news items by the application of plug-ins;
10) Should support for all level of sophisticated searching through the use of Boolean operators, positional operators, relational operators and fuzzy AND;
11) Should extend facilities for right truncated searching and case ignore searching;
12) Should allow field level and metadata based searching;
13) Should allow cross-collection searching;
14) Should support formulation of ranked query and Boolean query;
15) Should report number of hits and hyperlinks results with the target document;
16) Should provide extensive control over the display of news items;
17) Should help to process documents available in many formats such as html, pdf, text, rtf & ps;
18) Should help in processing both text and images;
19) Should accommodate collection covering a wide range of decided set of newspapers;
20) Should allow exporting of archives on CDROM in view of the poor Internet bandwidth in India;
21) Should be free of charge and provide ample scope in implementing the noble cause of 'right to information'.

Methodology

In order to achieve the objectives stated above, the following methodology has been adopted. The methodology in the design and development of VUNA primarily centres around three major issues:

- Architecture for web-integrated access mechanism;
- Selection of software;
- Building digital archiving environment;
- Operational steps.

Architecture

As far as architecture is concerned, the system incorporates three–tier access interface for news archive that includes any RDBMS as backend, CGI mechanism at the middle layer (with PERL, PHP or JSP) and HTML front end (SCHWARTZ & CHRISTIANSEN; LESK; OPEN SOURCE INITIATIVE (OSI) HOME PAGE).

Selection of software

The above mentioned architecture has been implemented by using open source solutions. Open source software (OSS) provide us the freedom to operate at system level. They allow customization of software through the modification of the source code (packed with binary) under General Public Library (GPL).

Green Stone Digital Library (GSDL), the widely known web integrated digital library software developed at the New Zealand Digital Library Project, by the University of Waikoto. GSDL, is having enormous literary warrant. This software is backed by the UNESCO for its implementation in developing countries. Automatic updation of the software, manpower training and other technical support to use this software are guaranteed by the UNESCO. Presently, a good number of libraries all over the world have resorted to this software, for building and dissemination of their digital collections, run in both UNIX and Windows environment. Due to these reasons, GSDL has been selected for indexing purpose.

The following open source software have been utilized at different layers for designing the system:

Backend layer:	MG (Managing Gigabyte) (KATHERINE)
Middle layer:	PERL and JSP
Indexing layer:	GSDL (OSS developed by University of Waikoto) (WITTEN & BODDIE)
Front end:	HTML and JSP
Web server:	Apache (APACHE SOFTWARE FOUNDATION)

Building digital archiving environment

The whole array of activities involved in the development of this digital archive may be divided into following groups:

Group I: Development of digital library environment

Building domain specific collection of newspaper items in digital form

Incorporation of DCMES and the domain specific newsML into the digital items,

scanned/keyed from the printed newspapers
Installation and configuration of Apache web server (Ver. 1.3.27) in Windows platform
Installation of PERL (Version 5.8.0. build 635 in Windows platform)
Installation of GSDL (Version 2.40 in Windows platform)

Group II: Development of web access mechanism
Configuring the system as a server and linking server (Apache) and digital collection (in GSDL) through modification of server configuration file (here *httpd.conf* file of Apache) to provide access to the digital collection in distributed information environment (MUKHOPADHYAY)

Group III: Organizing the digital collection through GSDL
Collection information; Source data; Configure collection; Build collection; View collection; Customization of user interface

Group IV: Development of off-line access mechanism
Exporting the digital library in CDROM for offline retrieval through stand alone PC by using a subset program of GSDL software

Operational steps
1) Selection of newspaper items
As this is a prototype system and intended for testing its ability as a model for building a news archive in India and Bangladesh to support study, research and application activities, only selected news items covering both text and image from leading ten Indian 'big newspapers' covering both English and Bengali have been incorporated. It is worth noting in this connection that our findings (MAITY) show that printed newspaper items are primarily of two types - text and images. Keeping this in consideration, items from a sample of ten newspapers, viz. Hindustan Times (Delhi), The Hindu (Chennai), The Times of India (Mumbai), The Statesman (Kolkata), The Telegraph (Kolkata); in English, and Ananda Bazar Patrika, Bartaman, Sangbad Pratidin, Aajkal and Ganashakti, in Bengali, published from West Bengal have been considered and collected.

2) Scanning/keying of selected news items
Using the flat-bed scanner (HP ScanJet 6300, 2400C) and OCR software (OmniPage Pro 10) text items are scanned while images are scanned using image processing software (HP ScanJet Pro Ver. 2.5). Besides some items are keyed in also.

3) Storage of items
Based on the nature and size of the items, they are stored in various formats. According to present state, texts usually take the formats like – pdf, ps, html, mht, rtf. So text items are stored in all these formats according to their specialities. Though images can be stored in various formats, keeping our purpose in mind to design a web-enabled system and speedier access, compressing capability and cost effectiveness, finally JPEG and GIF formats have been selected and image items are stored accordingly as needed.

4) HTML conversion

All the items stored thus have been converted to the html format to make the total collection searchable from the GNU Database Manager through web browsers like Netscape Navigator and Internet Explorer.

5) Metadata incorporation

Most widely available metadata schema, DCMES (DCMI HOMEPAGE) has been considered in this connection. Another domain specific metadata schema, NewsML (IPTC; REUTER' 2002a), has also been considered and the better flavour of both the schemas, has been incorporated. However for Bengali items, terms are used according to the prescriptions of a well known Bengali author's and editor's dictionary (*vide Bangla Lekhok O Sampadaker Avidhan: Ananda Bazar Patrika Byabohar Bidhi*. Kolkata: Ananda; 1994. 190p.). Figure-2 and Figure-3 are two sample pages showing metadata incorporation in English and Bengali newspaper items, respectively. The detailed discussion of metadata needs special attention and thus enumerated in following Section.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML><HEAD>
<meta name="Title" content="Musharraf may renege on pledge to quit as Army chief">
<meta name="Subject" content="Defence">
<meta name="Creator" content="PTI">
<meta name="Descriptor" content="Pakistan President Pervez Musharraf has said that
"the vast majority" of his countrymen want him in uniform as they fear that he would be
weakened without it.">
<meta name="Publisher" content="The Hindu">
<meta name="Contributor" content="Sandip Pathak">
<meta name="Date" content="20040918">
<meta name="Type" content="HTML">
<meta name="Format" content="text">
<meta name="Language" content="English">
<meta name="Coverage" content="Washington">
<meta name="NewspaperitemType" content="Top Stories">
<meta name="OfinterestTo" content="Senior Citizen">
</HEAD>
<BODY link=blue bgColor=#ffffff>
<TABLE width=770 border=0>
 <TBODY>
 <TR>
 <TD vAlign=top width=770>
 <TABLE height=40 width=770>
 <TBODY>
 <TR>
```

```
<TD align=left width=185>  </TD>
<TD align=middle width=400><IMG
src=”Musharraf may renege on pledge to quit as Army chief_files/hindu400.gif”
align=center border=0 valign=”baseline”> </TD>
<TD align=right width=185>  </TD></TR>
```

Figure-2: Display of metadata incorporation in news items (in English)

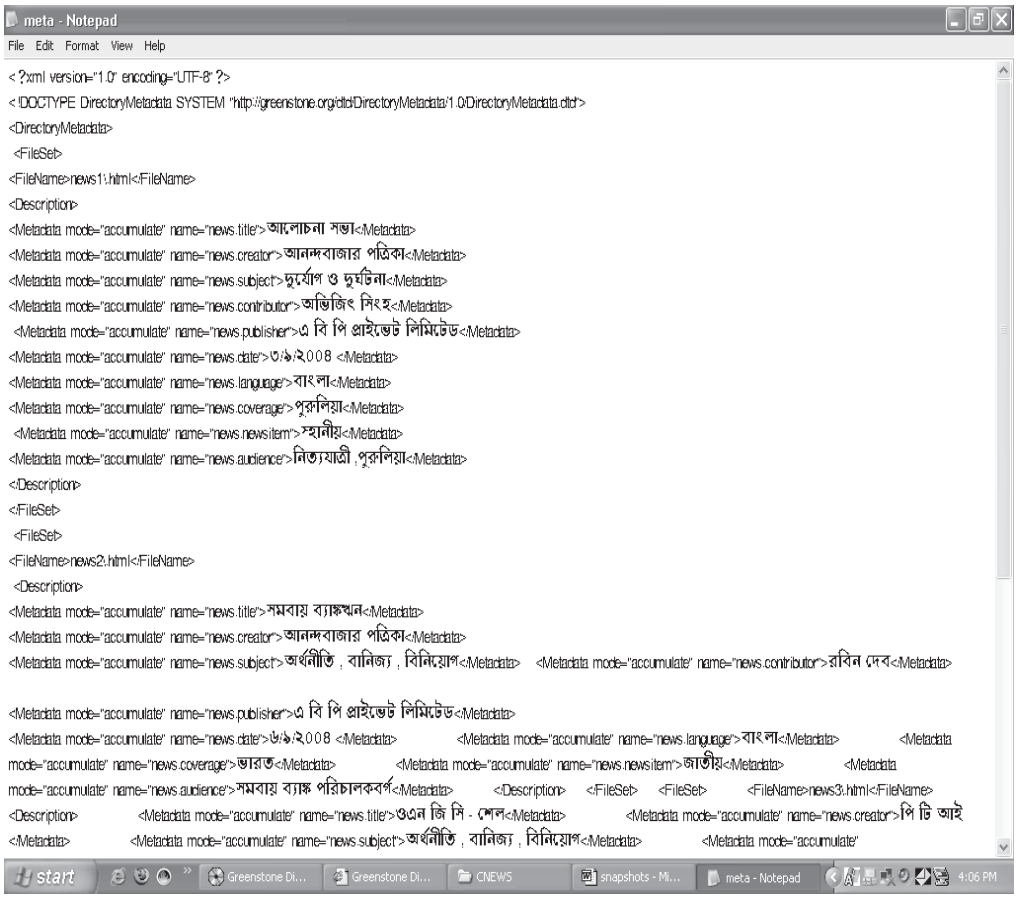

Figure-3: Display of metadata incorporation in news items (in Bengali)

6) Indexing

Indexing of the items using metadata schema is quite a straight forward method. But the crux lies in subject indexing. Details of subject indexing incorporating DCMES (DCMI HOMEPAGE), and Subject Reference System of newsML (REUTER, 2002b) from IPTC have been described in following Section.

7) Archiving

The 'collector' plays the role in archiving of the already saved items in the form of text and image. As this is a prototype system, and hence, no special provision has been made to archive multi-gigabyte data. This task also could be performed from command line, besides 'collector'.

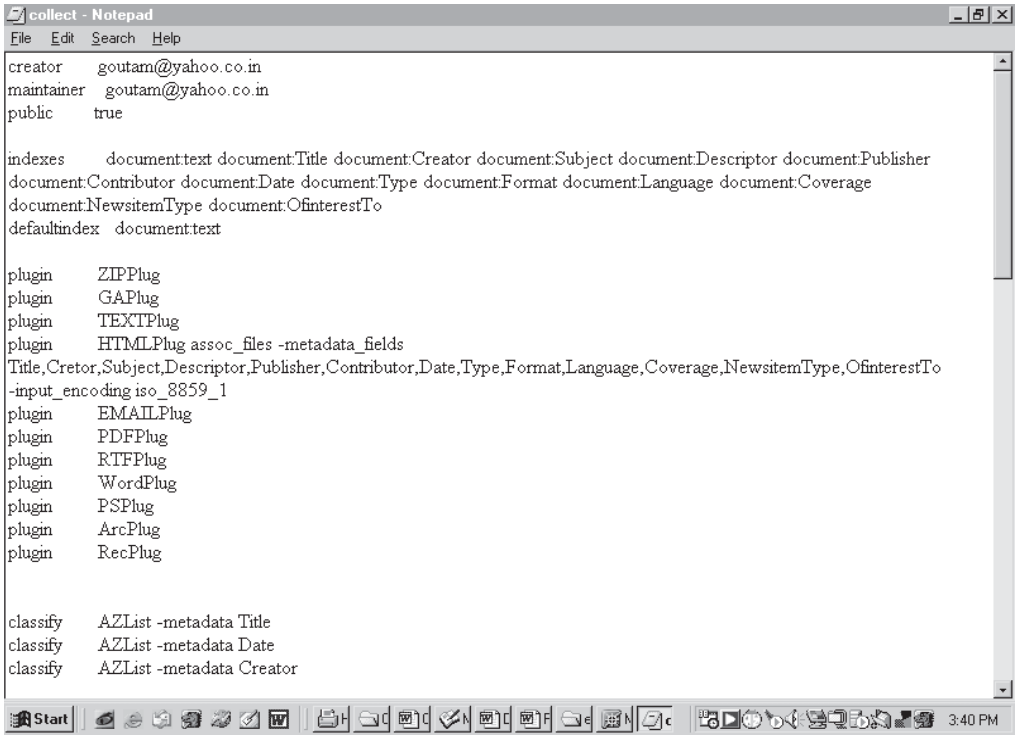

Figure-4: Configuring collection of the system

8) User interface design

User interface is one of the most important aspects of the system, considering that the users of the newspaper archives may require several options to satisfy their needs. It is designed keeping these factors in view. In the context of the multilingual, multi-cultured national scenario like India – the interface should be amenable for all kinds of users from various languages. Here the system has taken into consideration the Bengali language as to put the local emphasis along with the lingua franca English. Users may enter into the system by default being the member of a particular domain or users login and password could be provided for the authentication of users.

9) Accessing

Basically there are two modes of accessing to the newspaper archiving system - one is through web accessing, and another is through offline CD-ROM accessing. Both of them have their pros and cons. In India both the extremes are visible. In one hand – some industry and offices

got the information super high way, on the other hand as vast rural region with no/poor link to the civilized world. Keeping in mind both the features, this system has incorporated web-enabled mechanism, as well as offline-CD-ROM access to the news archive.

10) Use

Use of newspaper items includes browsing and searching. In case of browsing, the users usually do not come with any specific query, they lightly stroll on various news items. So, less options or access points are required. Search involves retrieving specified piece of information amongst the news items archived. Various search options alongwith examples have been enumerated in the following Sections.

Retrieval

In this connection retrieval includes browsing , searching and advanced searching, as well. Though the efficiency depends largely on the perfection of the implementation of metadata schema, generally. The retrieval of the news items will be done through:

Browsing

Browsing may be done through:

- Agency/Author
- Title
- Date
- Subject
- Types of newspaper item
- Target audience
- Place of origin, etc. (*vide* Figure-6,7)

Searching

Usually there are two sorts of specified search options:

Simple search : The following options are provided in case of simple search.

- Keyword
- Fielded search including author and title search
- Subject
- Free text
- Phrase search
- etc.

Advanced search: Advanced search is possible through the following options:

- Combined search through Boolean operators (AND, OR, NOT, XOR)
- Proximity operators, positional operators and relational operators
- Range search

Findings and Conclusion

The following (Figure-5) provides a view of the homepage of the system, wherefrom user(s) can have a brief idea of the system. The homepage provides links to English and Bengali newspaper collections. The users can select and enter into English or Bengali interface according to their requirements.

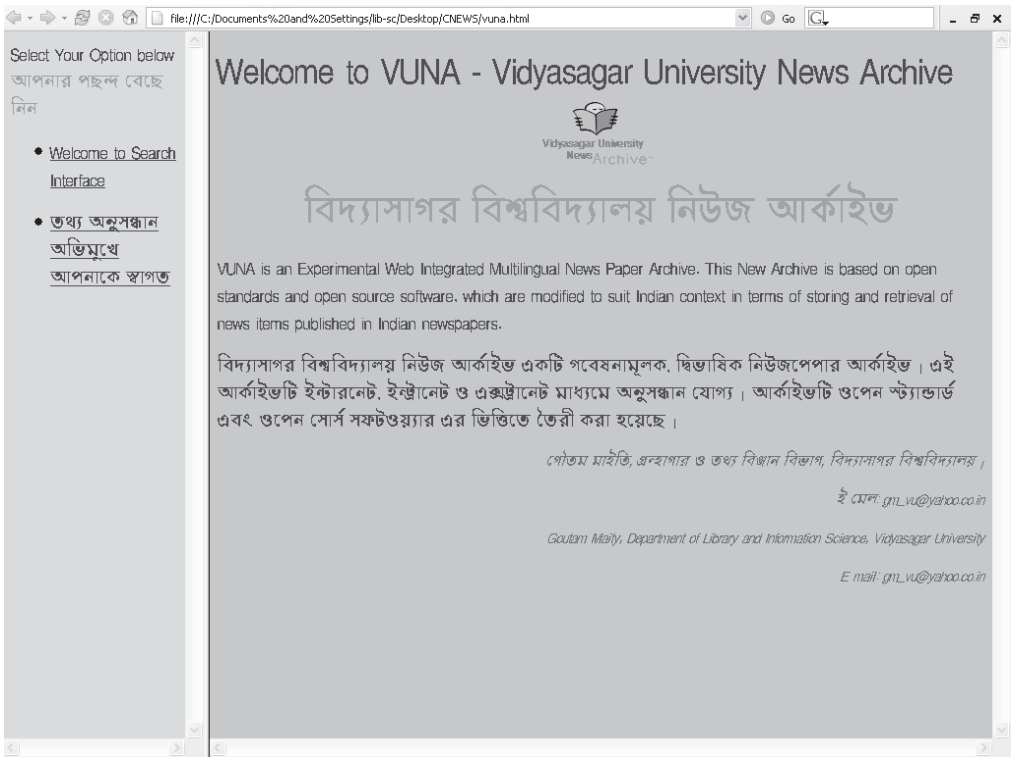

Figure-5: VUNA homepage

There are multiple ways to find information in this collection:

- search for particular words
- access publications by title
- access publications by date
- access publications by author (i.e., agency or person)
- access publications by topic/subject
- access publications by news agency
- access publications by publishing agency
- access publications by language
- access publications by types of news
- access publications by target group
- etc. (for details *vide* Figure- 6, 7)

The users can *search for particular word(s)* that appear in the text from the "search" page. This is the page that they can reach from the homepage through selection of option as to which collection – Bengali or English, they like to search and then clicking the same. They can also reach the appropriate collection from other pages by pressing the *search* button. The users can *access newspaper information by title* by pressing the *titles a-z* button. This brings up a list of items in alphabetic order. The users also can *access items by date* by pressing the *dates* button.

This brings up a list of all the items, sorted chronologically. The users may *access items by author* (i.e. agency or person) by pressing the *authors a-z* button. This brings up a list of items, sorted by author name. Apart from the above, they can also access items by several other options, as mentioned above and shown in Figure- 6, 7.

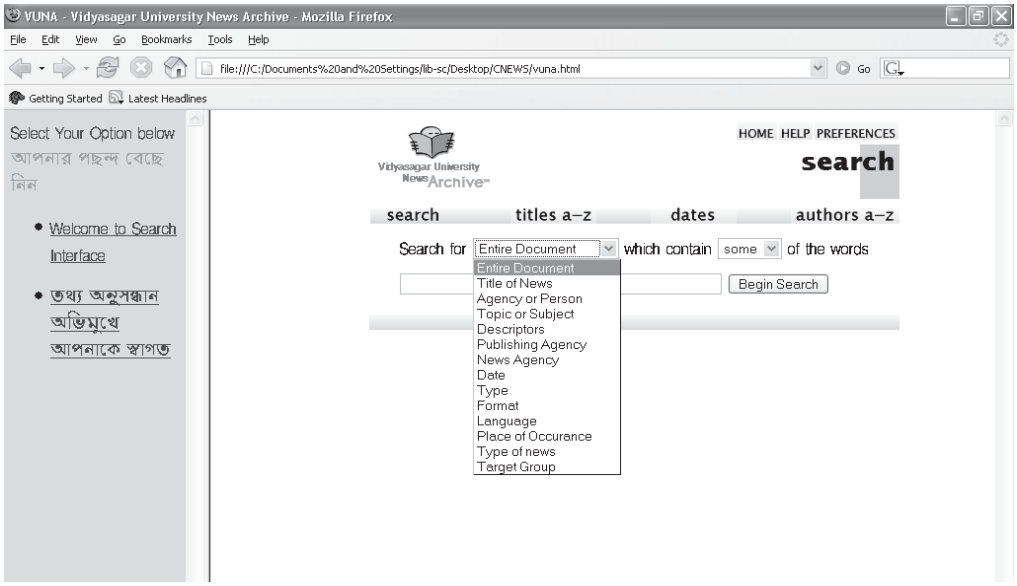

Figure-6: Search options in English interface

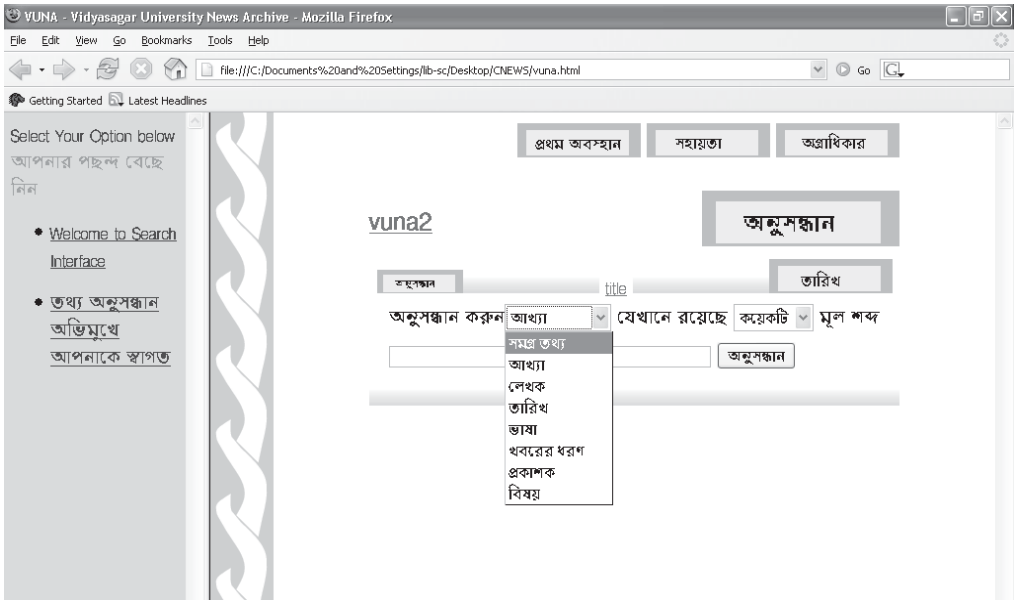

Figure-7: Search interface in Bengali

VUNA has a number of distinctive features. The reason behind the efficacy of VUNA is that it has been designed based on study of newspaper information generation, consumption alongwith an extensive study of existing newspaper archives on the web. However, it has used a web integrated architecture based on a set of open source software that provide freedom in customization and source code modification. Facilities for customization of user interface are available at the user end. Besides, VUNA is supported by international metadata harvesting protocol - OAI/PMH. Thus, it can allow seamless access over Internet, intranet and extranet. Considering the problems of huge volume of irrelevant items' retrieval in networked environment, VUNA has made a judicious use of general metadata schema- Dublin Core, and domain specific metadata schema- newsML, against test and verification to suit the Indian context, and thereby has enabled to achieve enough relevancy and precision in retrieval of newspaper items. Its policy regarding index exhaustivity and term specificity also helps in this regard. It allows browsing facilities through various metadata fields. Apart from simple search options including metadatabased searching, VUNA provides various advanced search facilities. As regards advanced search, it has enabled itself to provide facility for combined search through Boolean operators (AND, OR, NOT, XOR) truncation); searches using proximity, positional and relational operators; and range search. Combined phrase search option provided by VUNA may also be regarded as an advanced level search. These searches enable users to get more precision in retrieval. Due to its multilingual coverage, VUNA allows cross collection searching, which is a crying need in a multilingual and multicultural country like India. It can retrive full text news items both in Bengali and English. In addition, it helps the users to consult the search history. VUNA also provides extensive control over display of newspaper items. It supports inclusion/exclusion of any field in the display and provides scope of recording fields. It also provides online help. Format support (e.g. html, pdf, rtf, ps, etc.) in VUNA is adequate. Alongwith text items, it helps in processing of image items. VUNA is amenable for multiple languages, and it can accommodate collection covering a wide range of newspapers. Apart from that, it allows exporting archive on CDROM. It s cost effective and time efficient, and needs minimum users' efforts in an ultimatum. Searching is quite fast and ensures minimum response time because of organized storing of news items by the system. VUNA is totally based on open source framework and all the companion software (Apache, PERL, GSDL, web browser) are available from Internet at no cost. VUNA is totally based on open source framework and all the companion software (Apache, PERL, GSDL, web browser) are available from Internet at no cost. Considering the above, VUNA is totally a user friendly system.

As per LANCASTER, an information retrieval system can be evaluated by considering three basic issues:

- How well the system is satisfying its objectives, that is, how well it is satisfying the demand placed upon it;
- How effectively it is satisfying its objective; and finally
- Whether the system justifies its existence.

VUNA: a web integrated news information retrieval system that has been developed as a part of this research work may be analyzed in light of the abovementioned issues. VUNA is

basically designed to overcome the limitations of digital news retrieval systems, presently available in India. These information retrieval systems for news items are generally characterized by following features:

- use of proprietary software;
- poor user interface design;
- full text word-level indexing;
- no provision for field level search;
- large retrieval set and cross disciplinary semantic drift;
- no use of standard metadata sets;
- no provision for cross system export/ import;
- non-availability of on-line user help; and
- non-use of any standard subject reference system or vocabulary control device.

On the other hand, VUNA has been designed to ensure an efficient web-integrated news retrieval system, which can be used in libraries and information centres and also in news paper houses. In VUNA, digital news items can be exported to CDROM (along with a self-installation browser based program) to generate off-line information products. However, a comparative study of VUNA with available commercial and proprietary and other news retrieval systems in India and Bangladesh may be done under the following evaluation criteria, which are also in conformity with the parameters set by LANCASTER, mentioned above.

Table showing respective performance and failures of existing online newspaper archives in India and VUNA (i.e. Vidyasagar University News Archive)

Sl. No.	Features/ Check points	Existing online newspaper archives in India	Score (1 indicates presence; 0 indicates absence)	VUNA	Score (1 indicates presence; 0 indicates absence)
1	System architecture	No standard web-architecture is used to support MVC (Model-View-Control) framework	0	Supports standard three-tier web architecture to support MVC framework which in turn will support scalability	1
2	Use of standard metadata schema	Do not use any standard metadata schema (neither general-level nor domain specific	0	VUNA uses a combination of metadata elements based on Dublin Core (a general-level metadata schema) and NewsML (a specific-purpose metadata schema for newspaper items, developed by IPTC) with modification against test and verification to suit Indian purpose	1

3	Access to full text news items through web	These allow access to required full text news items against browse and earch	1	It allows full text access in various formats (like html, mht, pdf, etc.) against browse and search	1
4	Availability of on-line help	Not available	0	Available	1
5	Customization of user interface	No such facilities are available at the user end	0	VUNA allows extensive customization facilities at the user end	1
6	Export/Import	Export/Import is difficult in these systems, as these are not based on standard metadata schema and the collections are not OAI compliant	0	VUNA is completely OAI (Open Archive Initiative) compliant and based on internationally agreed metadata schema. Therefore, it ensures cross walk and interoperability at any level	1
7	Multilingual support	Not based on UNICODE standard thereby offers limited multilingual support	0	Fully compatible with UNICODE encoding (UTF-8 and UTF-16) and thereby offers support for all the scripts and language of the world	1
8	Generation of off-line information products	These are designed to support only online access through web	0	In view of the poor internet connectivity and bandwidth in India, VUNA is designed also to produce off-line CDROM product from on line accessible collection of news items	1
9	Freedom in source code modification	Not available	Not available	Available	1

10	Display of result	Fixed field display of retrieved result set is supported by these systems	0	VUNA supports inclusion / exclusion of any field in the display and also supports reordering of fields	1
11	Systems knowledge/skill requirement in implementation	These are based on easy to use, off the shelf software packages that demand minimum system knowledge/ skill at the time of implementation	1	VUNA requires extensive knowledge of web architecture (such as apache web server, CGI scripting, etc) and minimum knowledge about PERL programming language	0
12	Category based browsing	Offer limited facilities for category based browsing of news items	0	Offers extensive facilities for category based browsing such as subject, title contributor/author/ agency, news types, format, date, etc. VUNA also allows inclusion/ exclusion of categories on the basis of user feed back	1
13	Recall and precision	High recall and low precision	0	Relevant recall	
14	Collection coverage	Cover only newspaper(s) published by own establishment	0	As it's a prototype, it has been developed with ten newspapers but it suggests to implement it as a national newspaper information retrieval system	1
15	Field level searching	Not available	0	Available	1
16	Boolean searching (And, Or, Not)	Available	1	Available	
17	Use of relation-al and position-al operators	Not available	0	Fully available	1
18	Truncated searching	Available	1	Available	1

19	Proximity searching	Not available	0	Available	1
20	Phrase searching	Available	1	Available	1
21	Index exhaustivity and term specificity	Not available	0	Available	1
22	User effort	Require minimum searching skills for end user but the user has to formulate multiple search statements in order to retrieve relevant item(s). The users have to spent more physical-mental effort, time & cost	0	Simple search requires minimum searching skill. Although, advanced searching requires good amount of search skills. But in all respect it needs minimum effort from users' end	1
23	Response time in system	Intermediate to high response time required to carryout searches in web environment	0	Searching is quite fast and ensures minimum response time because of organized storing of news items by the system	1
24	Format support	Support only html format for storing digital news items	0	Supports a number of formats for storing news items (such as pdf, html, mht, ps, rtf, etc.)	1
25	Search history/ search profile storing	No such support	0	Supports are available	1
26	Availability	Some systems (e.g. content DM) are available commercially against price & annul maintenance charge	0	VUNA is totally based on open source framework and all the companion software (Apache, PERL, GSDL, web browser) are available from Internet at no cost	1

27	Cost factors for both system and the user	Systems involve high cost and users have to pay more	0	System involves comparatively low cost and the users have to pay comparatively less	1
28	Hardware/ Equipment	Require costly server-grade hardware & equipment Existing online Indian newspaper archives	0	Requires cheap PC-grade hardware as minimum support	1
	Total score		5	VUNA	27

(Score calculated in terms of binary code (i.e. 0 and 1))

The table, furnished above shows that the system - VUNA, designed and developed during this study, has achieved score of 27, where as other systems have gained a score of 05 only. Therefore, it may be claimed now that VUNA has reached its objective for which it has been designed and developed. Besides, it is also efficient and effective than other existing digital newspaper information retrieval systems in India .

Bibliographical references

- Apache Software Foundation. 1999-2004. Apache. Retrieved February 16, 2004, from http://www.apache.org
- Dcmi Homepage. 1995-2003. Dublin Core Metadata Initiative. Retrieved March 09, 2003, from http://www.dublincore.org/dc
- IPTC. 2003. International Press Telecommunications Council. Retrieved March 05, 2003, from http://www.iptc.org/iptc
- KATHERINE, D. 2001. MGPP: A Search Engine for XML Documents. Retrieved February 16, 2004, from http://www.greenstone.org/does/mgpp_ user.pdf
- Lancaster, F. W. 1979. Information Retrieval Systems: Characteristics, Testing and Evaluation. 2nd edition. New York: John Wiley; 1979.
- Lesk, M. 1997. Practical Digital Libraries: Books, Bytes and Bucks. San Fransisco: Morgan Kaufimanu; 1997. 336-346.
- Maity, G. 2006. Identification and Effectiveness of Factors in Designing a Newspaper Information Retrieval System (PhD Thesis Submitted to Vidyasagar University, Midnapore, WB).(Unpublished).
- Mukhopadhyay, P. S. 2004. Organization and Dissemination of Digital Objects through Web and CDROM: A Framework for Indian Libraries. In: Proceedings of the International

Conference on Digital Libraries; 2004 February 24-27; New Delhi: Volume 1. New Delhi; 2004. 470-478.

- Open Source Initiative (OSI) Homepage. 1998-2004. Open Source Software. Retrieved February 16, 2004, from http://www.opensource.org
- Reuter. 2002a. NewsML: The Markup Language for News Items. Retrieved March 05, 2004, from http://www.iptc.org/newsml/newsml.html
- Reuter. 2002b. Subject Reference System for News Items. Retrieved March 05, 2004, from http://www.iptc.org/newsml/newsml.html
- Schwartz, R. L.; Christiansen, T. 1990. Learning PERL. Cambridge: O'reilly; 1990. 267-270.
- Witten, I. H.; Boddie, S. 2001. The Greenstone Digital Library Software Manual. Retrieved September 15, 2003, from http://www.greenstone.org