Information Studies, Vol. 8, N o. 3, July 2002

Importance of Search and Retrieval in CD-ROM Full Text Publishing: Experiments Using PDF Documents and 'Nitya' Archival System

K.H. Hussain, R. Raman Nair and K. Raveendran Asari

As information technologies associated with CD, Web and Multi-media are making rapid advances; Electronic Publishing opens hitherto unknown areas in presenting, expressing and propagating ideas. Digital libraries are becoming a reality faster than librarians have dreamed or dreaded. No live library system can disregard these developments.

CD-ROMs in Digital Libraries

Few years back CD-ROM had been looked as a 'transient technology', in the hope that web technology would ultimately take over CD-ROMs. However, rise in number of commercial and technical CD-ROM titles shows that it has a big future. CD-ROM collections in the libraries are getting thicker day by day. Concept of 'paperless library' is now equated with CD-ROM Library. Its role in digital library will become more prominent with the spread of DVDs in coming years.

CD-ROM titles can be mainly categorized in to three:

1. Multimedia presentations: Audio and video CDs are produced in millions and becoming the most popular products of entertainments.
2. Indexing and abstracting services. E.g. 'Tree-CD' by Commonwealth Agricultural Bureau (CAB), AGRIS by FAO, etc.

3. Full text CDs: Though more and more digital libraries are created and made available through Internet, CD-ROM full text publishing is also flourishing. CD-ROMs are ideal for publishing reference works. Encyclopedia Britannica is now available in three CDs. Full texts of research reports and scientific papers on topics of academic interest are now a days collected and published in CDs. e.g. 'Solid Waste Management' by Environmental Protection Agency (EPA, USA).

Now a days any thing that is electronic / digital are put in CD, calling them 'Electronic Publishing'. Standards are yet to be evolved for electronic publishing, especially for CD-ROM publishing. Indexing and Abstracting services in CD are published with excellent search engines. 'WinSpirs', a retrieval package developed by Silver Platter is considered to be one of the finest search engines. It is used to produce important abstracting databases in CDs, such as CAB Abstracts, AGRICOLA, AGRIS, etc.

Full text CDs

Since CDs hold huge store of information, full text CDs resemble collections in special libraries. Strong similarities exist in between collection inside a library and a CD. While even a few hundred books are shelved in a library applying cataloguing and classification, thousands of document files are put in CD without applying proper documentation methods. Most of the CDs with full text are mere stack of thousands of files. This makes many full text CDs some thing like black- holes, where enormous quantity of information is roaming around inside with out finding a way to the out side world of information seekers.

Most of Full Text CDs are coming with a hypertext file, in which hundreds of titles are loosely classified under some categories. These titles are hyper linked with the original full text file, clicking upon which the full text is opened. In-depth free text search is impossible and users should be content with the categories provided. One has to browse all the titles under a category to get the needed one. If the title is inter-disciplinary, it may not be there in the category he is looking under. In some CDs of conference proceedings, thousands of papers are scattered under different sections. Looking for a particular title often becomes a formidable task.

In some CDs an index made of title words / key words are provided and hyper linked with the full text file. Here in-depth search is possible. However, the static nature of the index limits the search, since free combination of key words cannot be tried for retrieval.

Above cases underlines the need for alternate retrieval mechanisms in CD-ROM full text publishing. The dynamism of electronic media should fully be utilized in bringing out usable digital collections in CD by providing efficient free text search facilities. Such full text CDs are mass-produced by first developing a digital archive of document collections and creating a master CD.

Digital Archiving

Digital archiving constitutes one of the main components of digital libraries. It adds value and saves time while extending the hours of access. It reduces the need for proximity to information resources, but still emphasizes the quality of those resources. It is a library that can be individually customized and, ultimately, will be easy to use.

When Johannes Gutenberg of Germany invented the art of printing during the 1440s, the trend was to convert every manuscript into printed form. Microfilming technique helped to convert at least archival materials like old books, library catalogues, journals and newspapers into microform. What is prevalent at present is conversion of printed documents into digital form.

One can now easily scan thousands of printed pages and put in hard disk/CDs. More than thirty thousand word-processed pages can be stored in one CD. Although, several packages are in vogue for materializing digital archives, they lack efficient search mechanism. Devoid of a search mechanism, digital storage mediums will remain as dump places of information, just the same way as documents converted into microform. The full potential of digital archiving can be exploited only if selective access is made possible. ‘Nitya’ succeeds in achieving this goal.

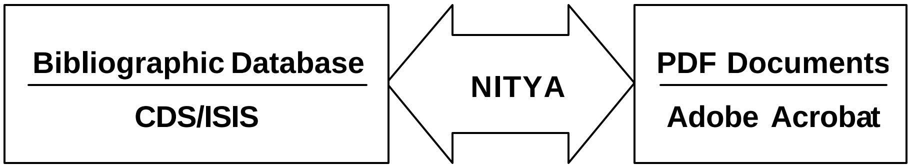
Nitya Archives

Nitya is an interface between CDS/ISIS by UNESCO and Adobe Acrobat. Nitya combines high-level text compaction technique and highly sophisticated free text search. The most outstanding aspect of Nitya is that any piece of information can be retrieved from a huge store of information within seconds.

Main objectives of Nitya are:

- It is meant for fast search and retrieval of digitized documents
- It is for creating stand alone Digital Archives.
- It is ideal for CD-ROM Full Text publishing.

Nitya has got two components. One is a collection of PDF documents created by Adobe Acrobat. Second is a database describing document elements using *CDS/ISIS.*



Adobe Acrobat

Acrobat is a brilliant e-Paper solution that converts scanned pages to PDF (Portable Document Format) documents. Conversion keeps all the page lay out features intact and compresses the file sizes to minimum. PDF files can be subjected to OCR.

Adobe Acrobat provides many hypertext facilities to make the documents `navigatable’. `Book marking’ is the most out standing one, which is made by headings of chapters and sections. This functions as a Table of Contents and clicking the bookmarks will open pages of respective chapters/ sections.

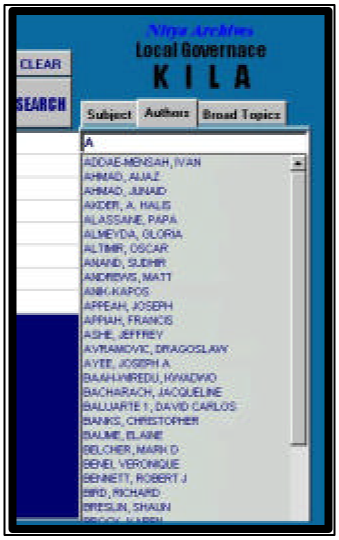
PDF documents have got many advantages over word-processed and scanned pages:

- Acrobat Reader is freely available and can be freely distributed to view PDF documents. Different 'pages views' are available and any part of the text/image can be magnified up to 1600 percentage.
- Different kinds of hyper links can be constructed easily.
- Acrobat retains lay out features of original pages and keeps the historicity of the document.

Features of Nitya

Nitya offers full functionality of windows. It only takes a few seconds to install. When it is opened it shows distinct areas for Dictionary, Query and Bibliographic exhibits. Dictionary contains searchable terms arranged under different categories such as subject, author, general topics, etc. Number of categories depends on the type and characteristics of the collection. (For example, categories of an archive of palm leaf manuscripts may be author, title, leaf no., first line, subject, etc, where as categories of an archive of theses may be the researcher, guide, departments, university, subject, etc). Search terms from the dictionary can be selected by a few alphabetical strokes. Clicking in the dictionary, terms are transferred to `Query' area. Queries are formulated using Boolean/proximity operators (OR, AND, NOT, *, NEAR). Queries are submitted to search and results are first exhibited in minimum details and then in full bibliographic details. When the 'Full Text' button is clicked, the full text is opened in Acrobat Reader and there after the document is navigated using book marks and other hyperlinks.

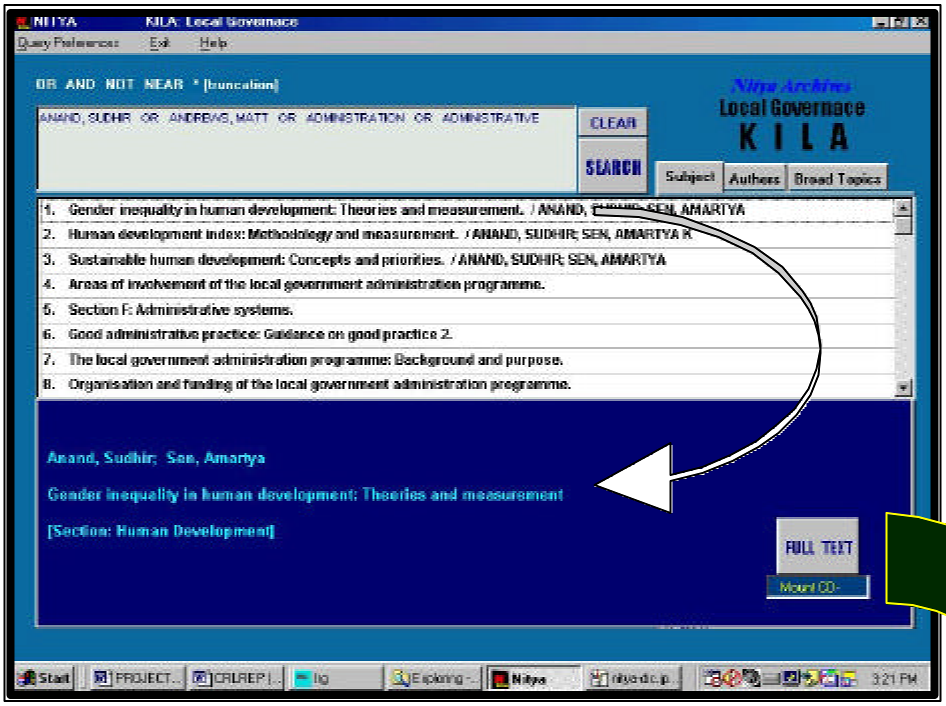
Dictionary and Categories



A

ADDAE-MENSAH, IVAN
AHMAD, AIJAZ
AHMAD, JUNAID
AKDER, A. HALIS
ALASSANE, PAPA
ALMEYDA, GLORIA
ALTIMIR, OSCAR
ANAND, SUDHIR
ANDREWS, MATT
ANIK-KAPOS
APPEAH, JOSEPH
APPIAH, FRANCIS
ASHE, JEFFREY
AVRAMOVIC, DRAGOSLAW
AYEE, JOSEPH A
BAAH-WIREDU, KWADWO
BACHARACH, JACQUELINE
BALUARTE 1, DAVID CARLOS
BANKS, CHRISTOPHER
BAUME, ELAINE
BELCHER, MARK D
BENEL VERONIQUE
BENNETT, ROBERT J
BIRD, RICHARD
BRESLIN, SHAUN

Search and Exhibits in Nitya



NITYA KILA: Local Governance
Query Preferences Edit Help
OR AND NOT NEAR * [truncation]
CLEAR
SEARCH
Nitya Archives
Local Governance
K I L A
Subject Authors Broad Topics
FULL TEXT
Mount CD-
Start PROJECT... CRLREP |... lig Exploring -... Nitya nityadoc.p... 3:21 PM

ANAND, SUDHIR OR ANDREWS, MATT OR ADMINISTRATION OR ADMINISTRATIVE

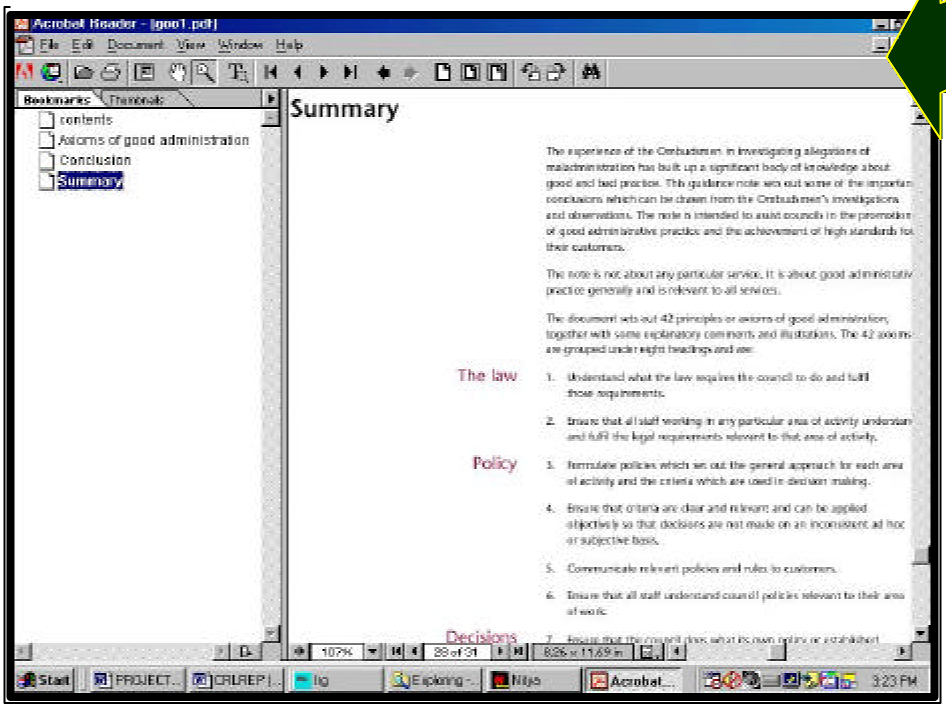
1. Gender inequality in human development: Theories and measurement. / ANAND, SUDHIR; SEN, AMARTYA
2. Human development index: Methodology and measurement. / ANAND, SUDHIR; SEN, AMARTYA K
3. Sustainable human development: Concepts and priorities. / ANAND, SUDHIR; SEN, AMARTYA
4. Areas of involvement of the local government administration programme.
5. Section F: Administrative systems.
6. Good administrative practice: Guidance on good practice 2.
7. The local government administration programme: Background and purpose.
8. Organisation and funding of the local government administration programme.

Anand, Sudhir; Sen, Amartya

Gender inequality in human development: Theories and measurement

[Section: Human Development]

Full Text in Acrobat Reader



Acrobat Reader - [goo1.pdf]
File Edit Document View Window Help
Bookmarks Thumbnails
contents
Axioms of good administration
Conclusion
Summary
107% 28 of 31 8.26 x 11.69 in
Start PROJECT... CRLREP |... lig Exploring -... Nitya Acrobat... 3:23 PM

Summary

The experience of the Ombudsmen in investigating allegations of maladministration has built up a significant body of knowledge about good and bad practice. This guidance note sets out some of the importan conclusions which can be drawn from the Ombudsmen's investigations and observations. The note is intended to assist councils in the promotion of good administrative practice and the achievement of high standards fo their customers.

The note is not about any particular service. It is about good administrativ practice generally and is relevant to all services.

The document sets out 42 principles or axioms of good administration, together with some explanatory comments and illustrations. The 42 axioms are grouped under eight headings and are:

The law

1. Understand what the law requires the council to do and fulfil those requirements.
2. Ensure that all staff working in any particular area of activity understan and fulfil the legal requirements relevant to that area of activity.

Policy

3. Formulate policies which set out the general approach for each area of activity and the criteria which are used in decision making.
4. Ensure that criteria are clear and relevant and can be applied objectively so that decisions are not made on an inconsistent ad hoc or subjective basis.
5. Communicate relevant policies and rules to customers.
6. Ensure that all staff understand council policies relevant to their area of work.

Decisions

7. Ensure that the council does what its own policy or established

Conclusion

Unlimited number of documents can be digitized, stored and retrieved using Nitya Archives. Old books, manuscripts in palm leaves, research reports, conference proceedings, parliament/legislative proceedings, theses/dissertations, government orders, journals, newspaper clippings, etc can be made in to digital archives and published in CD ROMs. Retrieval of a document is accomplished in seconds and thus Nitya solves many of the problems information access in CD-ROM Full Text Publishing.

Consultancy and training of 'Nitya' is undertaken by CIRD (Center for Informatics Research and Development, Thiruvananthapuram). CIRD is a group of professional librarians engaged in developing techniques for library automations. It aims transfer of technology to professional librarians, archivists and electronic publishers for digital archiving and CD-ROM publishing.

(For further details on Nitya and CIRD consultancy please visit the web site: cirdindia.org)

References

1. Breeding, M. *Does the web spell doom for CD and DVD?* Computers in Libraries. 19(10) 1999: 70-77.
2. Gopinath, *M.A. CD-ROM technology and its impact on Library and information science*. In: Devarajan, G. (ed.). Progress in information technology. New Delhi. Ess Ess Publications, 1996: 115-128.
3. Hasan, M; Muktiar Sing; Sharma, A.R. *CD-ROM: A powerful media for information packing, retrieval and dissemination*. In: Parthan, S. (ed.). Proceedings of the National Conference on Information Management in e-Libraries, 26-27 February 2002, IIT, Kharagpur. Allied Publishers Limited, New Delhi: 67-73.
4. Mishra, R. *CD_ROM: A medium of electronic publishing*. In: Sardana, J.L. (ed.) Library vision 2010: Indian libraries and librarianship in retrospect and prospect. Seminar papers, 45th All India Library Conference, 23-26 December 1999, Hisar, CCS Haryana Agricultural Library: 316-320.

5. Neena Sing. *CD Technologies and libraries.* In: Vyas, S.D. (ed.). Excellence in Information Technology. Dr. S.P. Sood Festschrift. Raj Publishing House, 2000: 124-131.

6. Raveendran Asari, K.; Hussain, K.H. and Raman Nair, P. *Nitya archives: Innovative blending of Techniques for selective Access to Information from Digitally Organized Text (SAIDOT)* In: Parthan, S. (ed.). Proceedings of the National Conference on Information Management in e-Libraries, 26-27 February 2002, IIT, Kharagpur. Allied Publishers Limited, New Delhi : 275-285.

7. Vijayakumar, J.K; Manju Das. *CD-ROM to DVD-ROM: A new era in electronic publishing of databases and multimedia reference source*. IASLIC Bulletin 45(2) 2000:49-54.

Authors

K.H. Hussain*, Documentation Officer, Kerala Forest Research Institute, Peechi, Kerala 68056, hussain@kfri.org*

R. Raman Nair*, Former Librarian, Kerala Agricultural University, ramannair_r@yahoo.com*

K. Raveendran Asari*, Head, Department of Medical Documentation, Mahatma Gandhi University, Kottayam, Kerala 68600, kravindran@vsnl.net.in*