# Architecture and building of Medical Digital Library at NIC [of India]: What exists and what is required for MeDLib@NIC?[1]

Mr. Sukhdev Singh, Mr. Surinder K Gaba and Ms. Naina Pandita.

Bibliographic Informatics Division,
(Indian Medlars Centre)
National Informatics Centre,
A - Block, CGO Complex, Lodhi Road,
New Delhi-110003. (India).

{sukhi, suri, naina}@hub.nic.in
Phone: 91-11-24362359

**Key Words:** Medical Digital Library; Digital Library Architecture; Building Digital Library; Web Standards; Information Integration.

## Abstract

ICMR-NIC Centre for Biomedical Information has developed various products that are available over Internet. These includes: i. UNcat (http://uncat.nic.in) - union catalogue of journal holdings of medical libraries of India;  ii. IndMED  (http://indmed.nic.in) - A bibliographic database of Indian biomedical journals and iii. medIND (http://medind.nic.in) - full texts of Indian biomedical journals being indexed in IndMED. Now, having these services, tools, databases and content in operation, the focus of future activities would be to integrate these "ingredients" both internally and externally to provide "single window digital access persistently". Here we propose an architecture under which each service, tool, database and content collection is an independent layer. These layers are the building blocks of Digital Library (DL) and can interoperate with each other due to either build-in or plug-in(ed) interoperability. They are accessible by their own interfaces as well as through Digital Library interface. In context of the proposed architecture, this article also takes stoke of what is available and what is required to build the digital library.

## Background

In 1986, National Informatics Centre (NIC) and Indian Council of Medical Research (ICMR) jointly set up a centre called ICMR-NIC Centre for Biomedical Information. This Centre was recognized as the 17th International MEDLARS Centre in 1990 and now is well known as Indian Medlars Centre (IMC). The centre provides wide range of services to biomedical scientists and medical professionals. These includes - Information Retrieval Services from various sources

---

[1] Edited and abridged version of a paper presented at International Conference on Digital Library, New Delhi, 24-27 February, 2004, entitled "Architecture and building of Medical Digital Library at National Informatics Centre: What exists and what is required for MeDLib@NIC?"

available on NIC hosts, Internet and CD-ROMs; supported by full text document procurement from NLM's DOCLINE and referral services from NIC's Union Catalogue. The content creation and aggregation activities include indigenous databases and full text journals. Evaluation and collection of Internet Resources and provision of chat room to end-users.  Training and user awareness programmes are conducted to enable users for better utilization of these content and services.

Last few years have seen IMC emerging as a content creator and aggregator. Taking advantage of penetration and popularity of Internet among user community, it has developed three major products that are available over Internet. These are:

i.      UNcat (http://uncat.nic.in)  - union catalogue of journal holdings of medical libraries of India;

ii.      IndMED (http://indmed.nic.in)  - A bibliographic database of Indian biomedical journals and

iii.      medIND (http://medind.nic.in)  - full texts of Indian biomedical journals being indexed in IndMED.
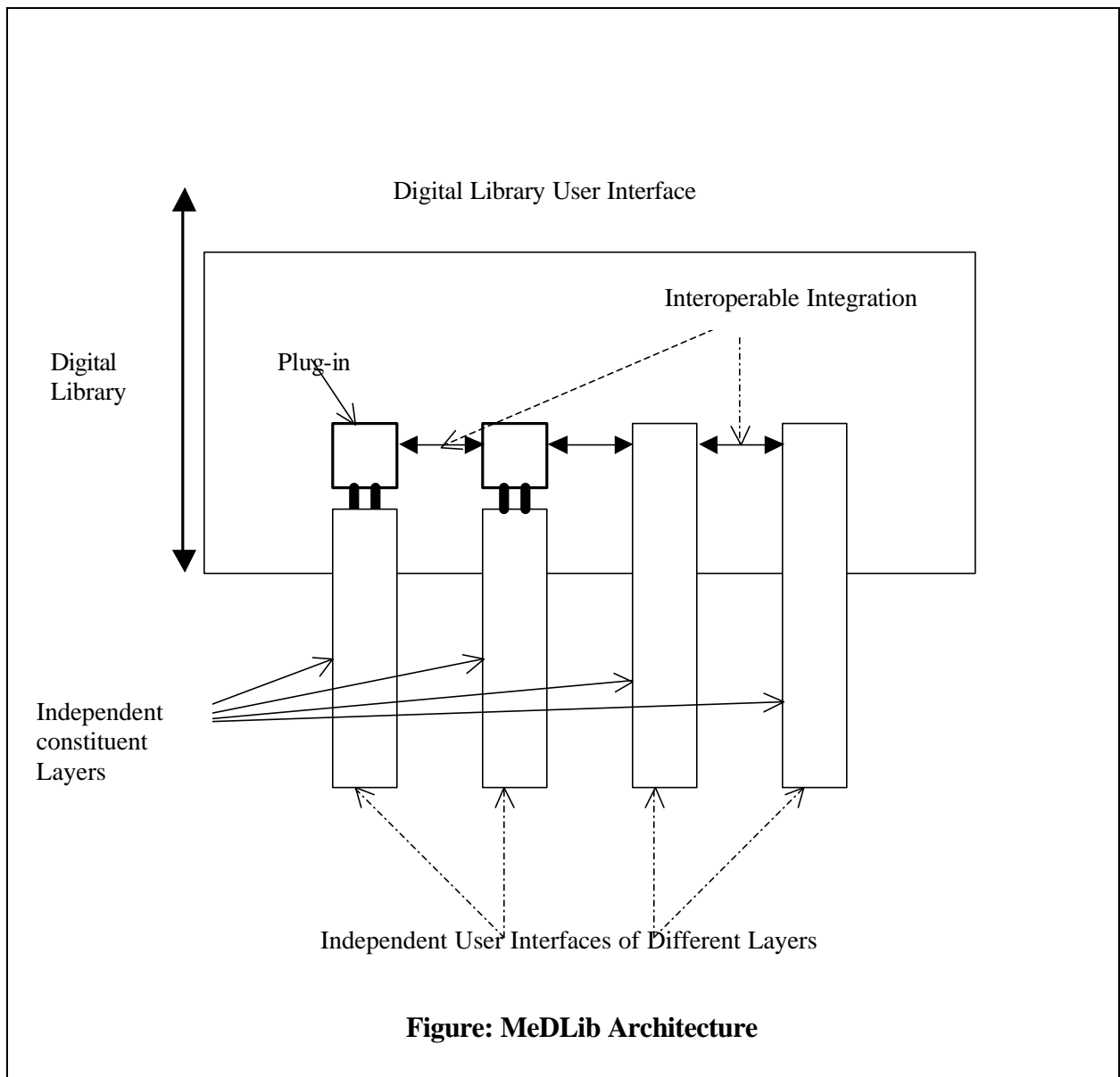
Apart from these major products many other services and content are available such as Live@IndMED Chat Room (http://chat.nic.in), collection of useful Internet sites and medical journals on Internet etc. In future, emphasis will be placed on not only such digital content and service collections but also on providing a digital environment for knowledge workers. This environment would aim in long run "to bring together collections, services and people in support of the full life cycle of creation, dissemination, use and preservation of data, information, and knowledge." (Duguid 1997).  A digital library would provide such an environment. Therefore future activities would be focused to build up a Medical Digital Library at NIC, that is, MeDLib@NIC. This digital library would capitalize on the existing services, tools, databases and content.  These would be integrated internally i.e. with one another as well as externally with those that are external to NIC Domain. The aim of this integration would be to provide single window digital access to collections of content and service in a persistent manner.

## MeDLib Architecture

For MeDLib@NIC, we propose an  architecture that we refer as "Vertically Stacked Layered Architecture". Here keyword "Vertically Stacked" has been added to signify two things. Firstly it is different from the "Layered Architecture" (Shaw and Garlan 1996) where each layer represents a different abstraction layer in the system as in the case of Open Systems Interconnection Basic Reference Model (ISO 1994). Secondly it means that the integration is internal to the system and there is no requirement for the constituent layers to be interoperable outside the system. Moreover integration is not intended to be visible beyond the system (here Digital Library) unless explicitly exposed. Under this architecture, each service, tool, database and content collection, is an independent layer. These layers are the building blocks of the Digital Library (DL). Layers can interoperate with each other due to either build-in or plug-in(ed) interoperability. Once interoperability is achieved, these are integrated into Digital Library. They are accessible by their own interfaces as well as through Digital Library interface. Thus here the constituent layers, under this architecture, are autonomous and heterogeneous services, tools, databases and content collections that have their own independent existence and may not be even aware of the integration. The interoperability among constituent layers is not a prerequisite because of the assumption that wrappers can be plugged-in for interoperability within the Digital Library.

MeDLib@NIC would consist of both internal and external layers. Internal layers refer to those services, tools, databases and content, which are under "direct control" or "influence" of NIC. Here layers under "direct control" means services, tools, databases and content collections that are made available to users by NIC. These includes UNcat, IndMED, medIND and others like chat room etc. Layers under "influence" means services, tools, databases and content collections provided by other agencies but following NIC's recommended standards, practices and protocols. Like Indian biomedical journals having their own independent web sites agreeing to follow NIC recommended standards, practices and protocols. While external layers refer to those publicly accessible, free or fee based services, tools, databases and content, which are beyond the control and influence of NIC. For fee-based layers an internal layer would take care of rights management. The Architecture would have the flexibility to accommodate the future internal and externals layers.

The architecture for MeDLib@NIC proposed here is at higher level of abstraction and does not yet takes into consideration any particular computational framework. Which would be mixture of hardware platforms, operating systems, protocols and applications. It would be difficult to restrict to any specific computational framework, as the constituent layers especially the external layers would be heterogeneous in nature and expected to be owned and controlled by different autonomous agencies.

Digital Library User Interface

Interoperable Integration

Plug-in

Digital
Library

Independent
constituent
Layers

Independent User Interfaces of Different Layers

**Figure: MeDLib Architecture**

Different strategies may be involved for integrating different layers. Strategies for the internal layers would be different from external layers. Build-in interoperability can be provided for internal layers while it would be difficult in case of external layers. For internal layers, the emerging frameworks of "Semantic Web" and "Web Services" could be considered for providing build-in interoperability.

**Semantic Web** is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. (Berners-Lee, Hendler and Lasilla 2001). The Semantic Web is the abstract representation of data on the World Wide Web, based on the Resource Description Framework (RDF) standards and other standards to be defined. It is being developed by the W3C, in collaboration with a large number of researchers and industrial

partners (W3C 2003). Semantic Knowledge Representation project of NLM is a practical example of semantic web application for accessing biomedical information (Rindflesch and Aronson 2002). **Web Services** is a technology that allows applications to communicate with each other in a platform- and programming language-independent manner. A Web service is a software interface that describes a collection of operations that can be accessed over the network through standardized XML messaging. It uses protocols based on the XML language to describe an operation to execute or data to exchange with another Web service (IBM 2003). Web services are invoked over the Internet by means of industry-standard protocols including SOAP; XML; and Universal Description, Discovery, and Integration (UDDI). They are defined through public standards organizations such as the World Wide Web Consortium. SOAP is an XML-based messaging technology standardized by the W3C, which specifies all the necessary rules for locating Web services, integrating them into applications, and communicating between them. UDDI is a public registry, offered at no cost, where one can publish and inquire about Web services. Web services technology has the ability to deliver integrated and interoperable solutions (Microsoft 2003).

While Semantic Web activity is dominated by research-oriented web community members Web Services is mainly backed by commercial interests including major plays like Microsoft, IBM and BEA. In the past there was a perception that Semantic Web and Web Services initiatives are in opposite directions. However Berners-Lee (2003) in his keynote address at the Twelfth International World Wide Web Conference has explained that these do not compete but can work together. Semantic Web can be used for data integration while Web Services can do the program integration. Further discovery mechanisms such as UDDI and WSDL are ideally placed to be implemented using semantic web technology; RDF could be sent as a SOAP payload, remote RDF query and update should use SOAP; semantic web business rules engines could interact using SOAP (Dumbill 2003). DARPA Agent Markup Language for Services (DAML-S) also demonstrates that Semantic Web and Web Services can be bridged together (Paolucci and Sycara 2003).

In case of external layers, the interoperability has to be plugged-in by various wrappers, plug-in(s) and agents. It would be the challenging aspect of building MeDLib@NIC. For layers representing databases, "Mediators" (Wiederhold 1992) can be used to retrieve data from underlying sources to answer the query. Other approaches are also available for the same (Jakobovits 1997). Web Wrapper agents can be deployed for layers representing web information sources (Chang *et al*. 2003). Such wrappers and Plug-ins would act as bridges between the native protocols and standards of external layers and common protocols and standards of Digital Library. This would ensure interoperability among all the layers.


**Building MeDLib@NIC**

According to our proposed architecture, MeDLib@NIC would be an integrated collection of autonomous systems (layers). The building plan would therefore emphasize on development and utilization of these layers. For layers under control and influence of NIC, i.e. internal layers, build-in interoperability is possible. The building plan would therefore include evaluation of existing layers for interoperability; modification to ensure interoperability if required and development of new interoperable layers. However build-in interoperability for systems not under control cannot be ensured. Here the plan would be to evaluate them for interoperability on regular basis and develop plug-in to utilize them in MeDLib@NIC. The building plan would also include development of at least two additional layers. One for rights management and other for ensuring persistency of the content and services.

The building of MeDLib@NIC would be based on sound principles. It would *expect changes*, will have prior *knowledge of the content* characteristics, *right persons* would be involved, emphasize on *designing useable systems*, strive for *open access* while still *respecting data rights*, *automate whatever is possible*, will *adopt and adhere to standards*, *ensure quality* and would be *concerned about persistence* (McGray and Gallagher 2001). Its processes would fellow best practices and standards. Where NIC does not have its own standards, it would fellow those endorsed by other agencies of repute including those endorsed by **Digital Library Federation** (2003). The practices and standards adopted would be endorsed by NIC for its partners to fellow.


## What exists and what is required?

Here we would review the existing layers along with preparedness and requirements for these to be integrated in the proposed Digital Library.

i. **IndMED**: It is a bibliographic database produced by IMC and available free of cost at http://indmed.nic.in. It indexes 77 Indian biomedical journals from 1986. IndMED provides much needed exposure to biomedical research published in India and acts an important resource to biomedical information seekers. It can be searched through WEB Interface using one of the three query forms (Singh 2001). The searching can be free text or restricted to various fields. Results can be displayed and redisplayed in required details. Hyperlinks to the source journals' homepages, if available, appear as part of reference information (Singh *et al* 2003). Links to full text articles in medIND are available where applicable.

An experimental grade semantic web of IndMED is available at http//medind.nic.in/imvw/. This is actually a by-product of another experiment (Singh 2002). Here each bibliographic record is published as an html file. Each of these files contains metadata in the html META TAG. The elements of this metadata are in accordance to Dublin Core specification. These files have been interlinks in tree fashion. This provides visibility to otherwise "invisible web" IndMED. It can be turned into full functional semantic web. There is further scope for improvements in processes related to updating of database. Presently the updating is manual. Indexers record data along with MeSH keywords in prescribed sheets. These sheets are later fed in the database by Data-Entry Operators. The same is then verified and edited by the editors. The process does not take advantage of online presence of some journals. Software agents could be deployed to harvest data for new issues of these online journals. This task could be made efficient by influencing these online journals to expose metadata in prescribed standard format. A semantic web application utilizing ontology based on MeSH or UMLS (Rindflesch and Aronson 2002) could then suggest keywords for each reference. This data after human intervention could go in the database saving significant time and effort. IndMED also needs to be exposed as web services so that it enables other systems to query and retrieve data.

ii. **medIND**: It is a project to provide access to fulltext articles of Indian biomedical journals indexed in IndMED. Indian journals indexed in MEDLINE would also be included in future. Presently there are 22 journals which available on medIND at http://medind.nic.in. MOUs have been signed with these participating journals to provide full text of articles. This understanding includes persistency of the articles made available on medIND. Many journals send their issues as soft copies. These are treated as master copies and are archived. Where softcopies are not available, hardcopies are scanned and archived as TIFF files. To provide access to users, PDF files are derived from these archived files. Browsing access is provided to these PDF files by html files providing links in Journal->Issue->Content Page->Article hierarchy. Search access is provided by

placing indicators in IndMED database which acts here as metadatabase for medIND. IndMED generates URL for the articles based on a file naming convention. The design of medIND is Internet Search Engine friendly. Search engines like Google can spider it and indirectly provide Internet wide full text searching capability. However there is scope for further improvements. The PDF files now exposed to search engines are devoid of metadata that is available in IndMED. Moreover the binding between IndMED and medIND is done manually. Agents may be developed and deployed to automate this binding and verify it periodically. Here the web services and semantic web can play crucial role.

iii. **UNcat**: It is a union catalogue of journal holdings of about 188 Indian medical libraries. It has been in existence for more then a decade. Its character based telnet interface was replaced by web-based interface few years back. UNcat is available at http://uncat.nic.in. Presently it is intended to be used by end users to locate nearest library holding a particular journal of interest. However, it has the potential of becoming a vital instrument in any initiative of medical libraries' cooperation. Exposing it as a web services may be considered. Then it may consume data from participating libraries' automation software and feed them with consolidated data in return. It can also be integrated with external layers like PubMed to indicate location of journals included in query results.

**iv. Information Sources Directory:** The existing listing of useful Internet information sources for health care professionals, researchers, consumers and medical librarians has the potential to be developed as information sources directory. This directory would include evaluated, rated and ranked entries describing each source in directory's database. Annotation along with location for each would be prepared and verified periodically. The directory would be exposed as semantic web to enable agents to consume data and produce subject guides giving overview of the sources available along with references to cited sources.

**v. Environment for Peer Interaction:** MeDLib@NIC will be in a position to provide a digital interaction medium to its users. This would provide a suitable environment to them for information sharing and creation. Interaction medium like the existing Live@IndMED chat room at http://chat.nic.in could be strengthened with additional services like Bulletin Boards, Blogs, List Servers and Virtual conferencing. Peer interaction could also produce intellectual content using "Commons-based Peer Production" (Benkler 2002) phenomenon. This phenomenon has been demonstrated to produce digital library content (Krowne 2003).

**vi. Self-Archiving Publishing Environment:** MeDLib@NIC would need to actively promote and provide suitable solutions and services to facilitate direct publishing of open access content by authors directly.

**vii. Multimedia and non-conventional content:** MeDLib@NIC would need to add multimedia contents and data into its scope. It could have repositories of Photographs, Audio-Video recordings and Medical diagnostic outputs etc. Its integration with Hospital Patient Management Systems could be possible. This may help MeDLib@NIC to become information repository for Medical Informatics applications.


## Developmental Plan

The development of MeDLib@NIC would be gradual and would be in three phases. Phase–I would capitalize on integrating the existing three core databases and content i.e. UNcat, IndMED and medIND; integration of IndMED and MEDLINE; strengthening of non-core activities like Chat Room and Internet resources collection; exposing IndMED data through different metadata protocols to have a semantic web of IndMED. Phase–II would focus in taking leadership for

promoting adoption of best metadata practices by Indian biomedical journals and their active involvement in IndMED and medIND activities; Integration of IndMED with non-medIND domain journals; development of electronic publishing solutions like E-Prints directed at the authors to encourage them to publish on MeDLib@NIC domain directly. Phase–III would strive for integration of UNcat and MEDLINE; working with Indian non-NIC efforts for access to biomedical information and information exchange including hybrid and conventional libraries.

**Conclusion**

We have proposed architecture for Digital Library. According to the proposed architecture, constituents of the Digital Library are autonomous systems and represent its layers. The layers, which are under the control or influence of owning organization, are referred as internal layers while others are referred as external layers. For internal layers the interoperability can be build-in, while it may have to be plug-in(ed) for external layers. In the context of the proposed architecture, we have taken stock of available and future layers along requirements needed to utilize them in the proposed Medical Digital Library. Such a Digital Library would provide single window digital access to content and services to biomedical researchers and medical professionals all over the world and could evolve as hub for cooperation among content producers, providers and other libraries.

# References

Benkler, Yochai. 2002
**Coase's Penguin, or Linux and the nature of the Firm**
*Yale Law Journal* **112**
http://www.benkler.org/CoasesPenguin.PDF (Accessed on 05-12-2003)

Berners-Lee, T. Hendler, J and Lasilla, O. 2001
**The Semantic Web**
*Scientific American* **284**(5): 34–43

Berners-Lee, T. 2003
**Web Services - Semantic Web [Keynote Address at Twelfth International World Wide Web conference, 20-24 May 2003, Budapest, HUNGARY]**
http://www.w3.org/2003/Talks/0521-www-keynote-tbl/ (Accessed on 05-12-2003)

Chang, Chia-Hui; Siek, Harianto; Lu, Jiann-Jyh; Chiou, Jen-Jie and Hsu, Chun-Nan. 2003
**Reconfigurable Web Wrapper Agents**
*IEEE Intelligent Systems* **2003** (September-October): 34–40

Digital Library Federation. 2003
**Digital Library Standards and Practices**
http://www.diglib.org/standards.htm (Accessed on 08-12-2003)

Duguid, Paul. (1997)
**Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments**
http://www.si.umich.edu/SantaFe/ (Accessed on 05-12-2003)

Dumbill, Edd. 2003
**Semantic Web and Web Services can live together, says Berners-Lee**
http://www.xmlhack.com/read.php?item=1978  (Accessed on 05-12-2003)

IBM. 2003
**New to Web Services**
http://www-106.ibm.com/developerworks/webservices/newto/  (Accessed on 05-12-2003)

ISO. 1994
*ISO/IEC 7491-1:1994, Open Systems Interconnection Basic Reference Model: The Basic Mode*
International Organization for Standardization.

Jakobovits, Rex. 1997
**Integrating Autonomous Heterogeneous Information Sources**
http://citeseer.nj.nec.com/jakobovits97integrating.html  (Accessed on 05-12-2003)

Krowne, Aaron.2003
**Building a Digital Library the Commons-based Peer Production Way**
*D-Lib Magazine* **9**(10)
http://www.dlib.org/dlib/october03/krowne/10krowne.html (Accessed on 05-12-2003)

McCray, Alexa T and Gallagher, Marie E. 2001
**Principles for Digital Library Development**
*Communications of the ACM* **44**(5): 49–54

Microsoft. 2003
**What are Web Services?**
http://www.microsoft.com/net/basics/webservices.asp (Accessed on 04-12-2003)

National Informatics Centre. 2003
**NICNET**
http://home.nic.in/htm/nicnet.htm (Accessed on 11-11-2003)

Paolucci, Massimo and Sycara, Katia. 2003
**Autonomous Semantic Web Services**
*IEEE Internet Computing* 2003(September-October): 34-41

Rindflesch, Thomas C and Aronson, Alan R. 2002
**Semantic Processing for Enhanced Access to Biomedical Knowledge**
In *Real World Semantic Web Applications*, pp. 157–172, edited by Vipul Kashyap and Leon Shklar
Amsterdam: IOS. 195 pp.

Shaw M and Garlan D. 1996
*Software Architecture: Perspective on an emerging Discipline*
Prentice-Hall

Singh, Sukhdev. 2001
*Web Enabling a Bibliographic Database of Indian Biomedical Journals: IndMED*
[MS Dissertation]
Pilani:  Birla Institute of Technology and Science

Singh, Sukhdev. 2002
**Publishing a CDS/ISIS Database in GSDL [Preprint]**
http://dlist.sir.arizona.edu/archive/00000187/ (Accessed on 11-11-2003)

Singh, Sukhdev; Pandita, Naina; Gaba, Surender K and Agarwal, Rajesh. 2003
**IndMED: Indian biomedical research database developed at NIC - a case study**
In *Electronic Information Environment and Library Services: a Contemporary Paradigm*
Delhi: Indian Library Association.
[Forty Eighth All India Library Conference, Bangalore, January 2003, Indian Library Association]
Eprint version available at http://dlist.sir.arizona.edu/archive/00000237/


US National Library of Medicine. 1997
**Press Release: Free MEDLINE**
http://www.nlm.nih.gov/news/press_releases/free_medline.html (Accessed on 18-11-2003)


W3C. 1999
**Resource Description Framework (RDF) - Model and Syntax Specification: W3C Recommendation 22 February 1999**
http://www.w3.org/TR/REC-rdf-syntax/ (Accessed on 05-12-2003)


W3C. 2003
**Semantic Web**
http://www.w3.org/2001/sw/ (Accessed on 04-12-2003)


Wiederhold, G. 1992
**Mediators in the Architecture of Future Information Systems**
*IEEE Computer* **25**(3): 38–49