Recent Trends in Databases and New Technologies for providing Online Library Services to create "Intelligent Libraries"

Manoj Kumar K, Scientist D (CS)

Library services have gone through phenomenal changes in providing information services by adopting proven technologies for dissemination of information to the users. Due to the fundamental shift in accessing information, latest tools and technologies are developed for meeting the versatility of platform for sharing of information over Internet. Over past few decades, databases played major roles in storing and indexing information in a structured way. Relational databases became important software for storage and management of data in conventional libraries. Various commercial as well as open source database management software have been introduced for various purposes. These software are having structured architecture to query and retrieve information through standard SQL. The compulsion to handle voluminous and structures data such as audios and videos has given birth to unstructured query languages like NoSQL databases. Similarly for the user interfaces, simple HTML and CSS have been replaced with more sophisticated tools like JSP, JSON, XML, XHTML & HTML 5, J-Query, DOM, AJAX, HTML5, built-in Java/PHP framework, etc. Device and browser independent applications are in demand due to the penetration of smartphones, tablets and other hand-held wireless devices amongst masses. These kinds of latest devices are widely used by academic community and users also demand for the applications which can easily run on these devices. Android, Windows Phone, Palm WebOS, Blackberry etc. are commonly used by students and academic community who look for the library services to run on these devices. To reach services and applications to the all strata of the society, the services should be available in vernacular languages also. This paper discussed evolution of databases and library automation software, trends in technological changes in application development, innovation in providing services, semantic web, ontology, recent front-end tools for user interface design and strategies to be adopted to choose such trends and tools.

Evolution of Technologies in Libraries

In 1960s, on demand batch searching (offline) was used in MEDLARS systems for search service by using computer and later dialogue and chemical abstract services developed. Online services start with introduction of packet-switched data communications network. In 1970s more than 300 public access databases were available. This kinds of services like MEDLARS, MEDLINE, Dialogue, ORBIT, LEXIS etc. led to the inclusion of Online Public Access Catalogue (OPAC) in library automation packages.

Modernisation of libraries start with introduction of the automation of routine activities of library functions. Acquisition, technical processing of the collections in the library which includes cataloguing, classification and accessioning, circulation (issue, return, fine calculation and reservation), user management and administration, serial control management, budgeting,

stock verification, binding process, OPAC, etc. are major activities of any library. The technological evolution has also reached to a level where the fundamental way of managing the library activities should be changed. Use of electronic devices and computers has transformed manual activities of libraries into modernised way by using relevant software. SOUL, Koha, NewGenLib, ABCD, Evergreen, e-granthalaya, Alice for Windows, Libsys are some of the major software used in Indian libraries for automating the activities in new era. Internet and allied services have given birth to formulate new methodologies and technologies to disseminate the information into various forms and formats. Static web i.e. web 1.0 is introduced to provide information on a web page to the user and services like WebOPAC etc. are introduced a decade back in libraries. These services are accessed by the users not only with PCs but also with latest gadgets such as smartphones, tablets, phablets, iPads, etc.

Library Management Software using Latest Trends

The libraries are using the software mentioned above for its automation activities. The evolution of such software has adopted the tools and technologies available during the period of development and migrated to latest version depending on further development. WebOPAC is an implementation of web 1.0 technologies in library management software. Many of the software is redesigned to equip with web technologies and migrated to web based automated services. The client server model is shifted to web server/browser model on Internet platform.

Recent trends in technologies and gadgets are challengingly forcing the libraries to adopt latest communication tools for effective delivery of knowledge and information to the user in cost effective manner. Web 2.0 tools and technologies are playing a major role in libraries for the interactive model for the exchange of information with user satisfaction. The web 2.0 tools such as social networking, social bookmarking, blogs, weblogs, folksonomies, RSS feeds, wikis, mashups, widgets, on-demand video streaming, video sharing sites, hosted services, podcasting etc. are integrated with existing library technologies for providing hybrid services to the user in desired interactive way. Customised web services with its simple architecture is using the HTTP protocol with XML serialisation in conjunction with other web related standards for the representation of web resources using a uniform set of stateless operations (REST) in libraries now.

Web API (Application Programme Interface) is a new development in web service. Web enabled REST based communications are used as shown in figure 1. RESTful APIs is based on stateless architecture which do not always demand XML based web services protocols (SOAP & WSDL) to support their interface. Web APIs are required for exchanging data over Internet when two different software systems are used. The software system that requests data is called service requester and the system that provide data for the required system is

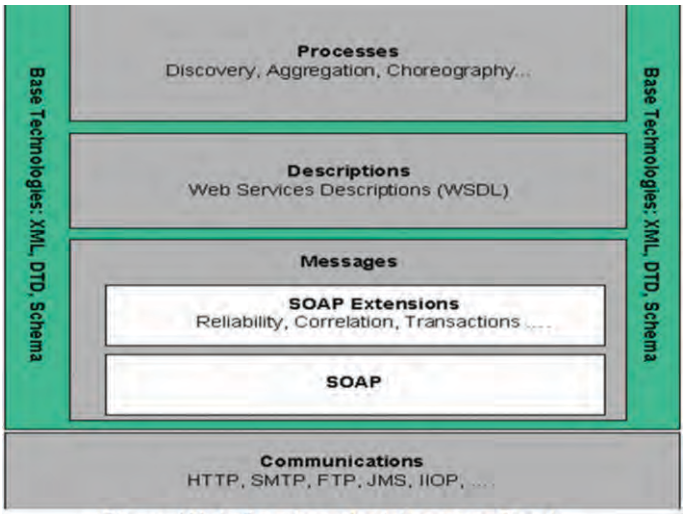

Fig. 1. Web Services Architecture Stack

called service provider. The service provider sends a WSDL file (Web Services Description Language) to a central directory called UDDI (Universal Description, Discovery and Integration). UDDI defines which software system should be contacted for the type of data requested. After finding out the software system, it uses a special protocol called SOAP (Simple Object Access Protocol) and sends structured data in an XML file using SOAP protocol. The XML file would be validated again by service requester using an XSD (XML Schema definition). Modern library management software is exploring the possibility of including this model of web services in its designs. The web services that used markup languages are:

- Representational State Transfer (REST)
- JSON-RPC/WSP (Java script Object Notation Web Service Protocol)
- WSDL
- WSCL (Conversional Language)
- WS-Metadata Exchange
- XML RPC (Remote Procedure Call)

Web services based technologies enable the libraries to host their services in cloud and flexibility of building the software as well as platform as service. The libraries are redefining the software with a new design pattern as Service-Oriented Architecture. This kind of service orientation is independent of any vendor, product or technology. Such adaptation of web services in libraries will change the conventional style of automation of libraries in future.

Building Digital Libraries

Along with automation of libraries, digitisation is also considered as one of the major task of modern libraries. Collaborative content creation is the mantra of web 2.0 technologies. Scanning and preserving the content is primary objective of the libraries in modern era. OCR based searchable content is required to be created for the effective search and indexing of the content in libraries. Academic materials, research resources, manuscripts, institutional documents, publications etc. are to be digitised for preservation and sharing.

In the current era, all documents are almost digitally born and digitisation is as simple as converting into open standard document such as searchable PDFs, TIFF, etc. Old documents and manuscripts are to be scanned digitally to create digital fingerprints. Scanning hundreds of thousands of bound volumes and books without damaging them is a challenging task. Latest scanners have ability for creating scanned books at a rate of 250 pages per minute. The slow manual scanning is replaced with technologically advanced high speed non-destructive scanning of the book in its original format by robotic book scanner. As the pages flip for scanning a pair of high definition cameras scans the pages and laser lights provide information about the page's 3D topography. 3D topography allows the robot to map the pages 3D deformation and using a real-time algorithm it restores the image to a flat view without curling as well as guarding against skipped pages. Thus, practically, 1.2 lakh pages can be scanned in a day. If the books are available as unbind volume with loose papers, flatbed scanners or automatic document feeder (ADF) can be used for scanning. Pages with a riffled edging or curving due to a non-flat binding can be difficult to scan using ADF. ADF is designed to scan pages of uniform shape and size, and variably sized or shaped pages can lead to improper scanning. Latest ADF scanners have the ability to scan in colour/greyscale/monochrome with 200-300 dpi at the rate of 80 pages per minutes or 160 images per minute. These kinds of scanners are widely used in universities for digitisation of theses. Digitised contents can be aggregated into a digital library or Institutional

Repositories (IRs) for the benefit of user based on policy of the institute. Copyright and IPR should be checked before making it available for the users.

Content Creation and Management of Repositories (IR and OR)

After creation of digital scanned PDFs, the repositories are to be created for searching the scanned documents. The structure of the digital content, communities, collection and metadata are to be created and assigned for each document. The conventional database has its own limitation for storage and retrieval of huge digitised achieves. There are many open source software which are available to address these issues with built-in architecture and data models. DSpace, Greenstone, e-Prints are such popular open source software which can be used for creation of Institutional repositories and digital libraries. Many of the software have web 2.0 built-in features for the dissemination of digitised content for the users. The metadata has to be created as per the international standards such as Dublin Core, MARC21 etc. Modified Dublin Core can be used, if the collections are specific to any document type. Customisation of the user interface may be done with the help of scripting language such as JSP, Java Script, PHP, Perl, etc. A database is also required for storing the content in an effective and systematic manner by auto indexing the content. There are several open source databases available which will be discussed in next section. Normally the design of any digital library software is divided into 3 layers i.e. storage layer, business logic layer and application layer. A sample architecture used in DSpace is given in Figure 2.

Open repositories can be created with handle IDs which is the unique and persistent identity for the digital objects as defined by CNRI's Digital Object Architecture. There are global handle registries for maintaining the handles and generally, handle services are being run by national libraries, national laboratories, universities, government agencies, research groups and federations. For example Shodhganga is having handle identifier at URI: http://hdl.handle.net/10603 which is registered with CNRI. The theses in the repositories are having unique identification number which is numerically coded i.e. http://hdl.handle.net/10603/17937 is ID assigned to one of the thesis hosted in Shodhganga.

Now a days, each researcher is identified by a unique global research identity which will help to create a transparent method of linking research activities and research output linked to these identifiers. ORCID ID, Google Scholar ID, Researcher ID, Scopus ID etc. are some of the popular researchers ids used globally. Libraries can promote their researcher to create IDs in any or all of these global registries. ORCID also provide APIs to support system to system communication and

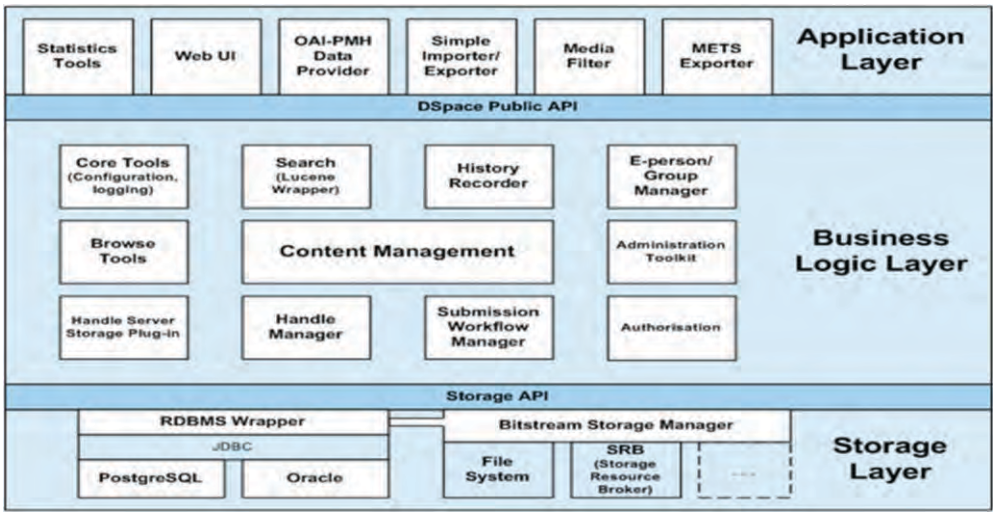

Fig. 2: Dspace Architecture

authentication under open source licence. Vidwan database of INFLIBNET Centre, which is national database of experts in various disciplines integrated with ORCID ID for publication details to populate in the user profile. Likewise RingGold is a Cetral registry of that maintain a unique ID for Institutions and organisations.

Management of Data and Storage

The demand to store voluminous data in libraries requires sophisticated database management systems (DBMS). There are more than 200 different database systems that can be use as backend storage systems. Traditionally relational DBMS are used to store data in a structured and organised manner. RDBMS is the basis for Structure Query Language (SQL) and other modern database systems like MSSQL, MySQL, PostgreSQL, Oracle, IBM DB2, Sybase ASE and Microsoft access. Data in RDBMS stored in database objects called 'tables'. Though RDBMS is adequate for data which is in simple tabular structure, it is difficult to model for complicated and complex data structure which is having multiple levels of nesting. This can be resolved by using Open Source Database NoSQL (Not Only SQL) which is used in Big Data and real time web applications.

NoSQL and New Data Models

NoSQL is a new approach to data management and database design which is useful for very large set of distributed data. Since NoSQL supports a wide range of technologies and architecture, it is preferred to solve the scalability and big data performance. NoSQL is widely used to manage huge unstructured data which are stored on multiple virtual servers in the cloud. The benefits of using NoSQL over relational databases are its capabilities to handle large volumes of structured, semi-structured and unstructured data and support object oriented programming and scale-out architecture.

The most popular NoSQL database is Apache Cassandra which was Facebook's proprietary database and MongoDB is also growing rapidly. The other NoSQL implementation includes SimpleDB, Google BigTable, Apache Hodoop, MapReduce, MemcacheDB and Voldmort. Some of the popular social networking sites which use NoSQL include Twitter, Linkedin, NetFlix, etc. NoSQL database types are document databases which can contain (i) different key value pairs, key-array

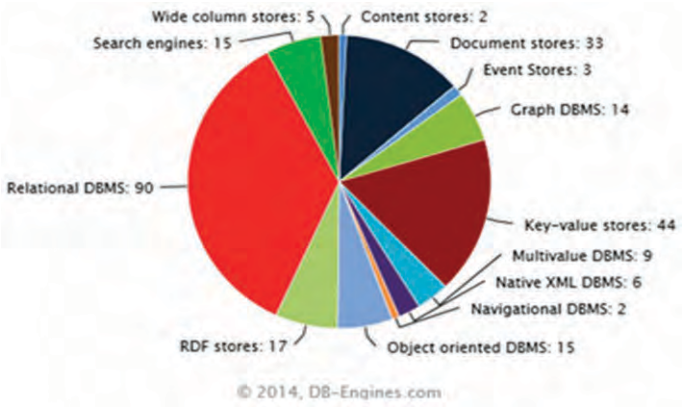

Fig. 3 : Ranking by Database Model

pairs or even nested documents, (ii) graph stores-to-store information about networks such as social connections which include Neo4J and HyperGraphDB (iii) key-value stores-to-store attribute names with its values, eg. Riak and Voldmort. (iv) wide-column stores such as Cassandra and HBase which are optimised for queries over large datasets and store columns together. The distribution of RDBMS as well as NoSQL Database types is shown in the pie-chart.

User Interface & its Development Tools

The storage issue is being address by modern open database like NoSQL and old RDBMS for databases.

Properly trained Database Base Administrators (DBAs) along with strict policy enforcement for maintenance of database is required. Stack overflow, SPOF (Single Point of Failure) which is potential risk posed by flow in the design, implementation and configuration of database systems and networks is to be planned during the design of the system where big data is used. The word 'Big Data' stands to describe massive volume of structured and unstructured data. The big data is not only used to refer to the volume of data but also the technology i.e. tools and processors that a library requires to handle the large amount of data and storage facilities. The data size may be petabytes-PB (1024TB) or exabytes (1024 PB) of data.

In such a situation, to deal with 'Big Data', distributed indexing, replication and load balanced querying, automated failover and recovery with fast search interfaces are required. These kinds of challenges are resolved by advanced search server with a REST like API as mentioned above. Then unstructured documents are indexed with latest tools such as JSON (JAVA Script Object Notation), XML, CSV or Binary over HTTP. Solr is an example of open source search platform from Apache Lucene project which include all the features mentioned above and also support hit highlighting, faceted search, dynamic clustering, database integration with NoSQL feature, rich document handling etc.

For programmers, customised interface can be made by using programming languages/tools such as PHP, Perl, cross-platform JavaScript Library like J-Query, DOM (Document Object Model), AJAX, HTML5 etc. AJAX is very popularly used as a interrelated web development techniques used on client side to create asynchronous web application. It is also a preferred script for exchanging data with server and updating a small portion of webpage without reloading the whole page. HTML 5 addresses the issues of scalability in HTML 4 and have introduced additional features to support semantics nature of web, innovative connectivity, offline and storage, Open web multimedia, diverse range of presentation options for 2D/3D graphics and its effects, allowing for the usage of various input and output devices and styling.

Ontology and Semantic Web

The introduction of web technologies and World Wide Web in 1990s has given innumerable opportunities for many services and platform for sharing of resources. New services like Email, FTP, Search Tools, and HTML for web, discussion groups, hosting of various services in a single server are introduced in the initial phase of Web 1.0. From the static nature of web document and read only web, the interactive and social web is introduced for collaborative content creation and user interaction. Read-Write-Modify-Distribute applications and contents became the heart of the Web 2.0 technologies. Search and retrieval of content was part of the visible web.

Retrieving the content from multiple sources and intelligently analysing the data is challenging task. Creation of knowledge from hyperlinked documents and semantically linking the useful and related documents is a research topic in the modern web. Semantic web (coined by Tim Berners-Lee) is collaborative movement by international standards bodies like World Wide Web Consortium (W3C) for converting the convert web which is dominated by unstructured and semi structured documents into a "meaningful web". As per W3C, semantic web provides common framework of semantic web stack which is built on W3C Resource Description Framework (RDF) to allow data to be shared and reused across applications, institutions and communities. The

technology enables computers and people to work in cooperation. Since digital library is an interconnected library of documents by creating hypermedia of links which retrieves data from distributed databases, syntactic web cannot give desired result to the focused users. Semantic web using the concept of human deductive reasoning and inference logic based on new technologies like RDF, Ontology Languages, XML etc. can produced meaningful result and help computers to perform automated information gathering and research. Ontological model of data and description is required for relating different pieces of data to another. Ontology is a standard way of modelling the ways different pieces of data relate to one another. This help machine to understand the data and relations by adding meaning to metadata. Thus knowledge can be represented as a set of concept with a domain and relationships between those concepts. Ontologies are nothing but explicit specification of shared conceptualisation. Modern well designed databases tends to deal with lots of relationship between different elements of data. Ontology vocabulary with RDF, XML, Unicode, URI with trust proof and logic (rules of modern Web) becomes the semantics web vision for knowledge retrieval. Web Ontology Language (OWL) extends RDF and RDFS (Schema) for interpretation of web content with the above technologies for mission interpretation. OWL-Lite, OWL-DL, OWL-Full or DAML (DARPA Agent Markup Language) along with reasoning tools such as Jena, RacerPro, FaCT++, HermiT, Pellet, etc. and ontology interface layer is used for retrieval of 'Web of Knowledge' from the deep or invisible web. Modern query languages like SPARQL, DL Query, nRQL, RDQL, OWL-QL, SWARL could be used for extracting exact meaningful content from the semantic web.

Strategies for Choosing Trends and Tools

Technologies are changing every day and new techniques / tools are introduced every day. Many of such technologies get obsolete after creating a small hype (eg. Pager). It is absolutely difficult to realise the impact of technology or tool in short time. Social networking also plays a major role in dissemination of information by connecting to respective users. The strategies are to be chosen by learning and observing the changes that the technology is making in society. It is suggested to follow the trends in following in order to choose tools:

1) Hardware trends
2) Networking trends
3) Development and changes in programming languages
4) Trends in web technology
5) User feedback about using a new tool
6) Percentage of growth of technology and tool (Google trend analysis can be used to find out) an example of trends of two NoSQL databases mongoDB and Cassandra given below

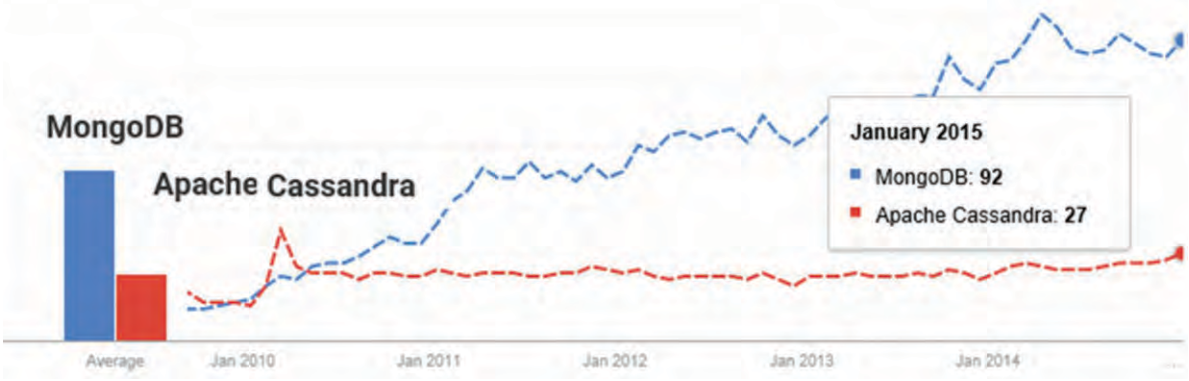

Fig. 4 : Analysis by Google Trend for two new Databases

Conclusion

Libraries are going under radical changes due to quick technological development. The online resources are created in libraries to provide hybrid services to the users. Online contents such as e-journals, e-books, e-scripts, e-questionnaires, videos, audios, animations, play major role in providing content to the users for academic needs. Conventional databases, programming languages, scripting tools are being replaced with advanced tools and software. Reliability and scalability need to be assured for the progression growth of contents with multi formatted e-content. The three layer logic for designing a system replaced with entire model by liberation of the data from the logical and interface design. Speed and accessibility of content, where bandwidth is not a barrier, is directly proportional to the technologies used behind solution used for providing online services. Traditional RDBMS like MS SQL, ORACLE, MySQL, PostgreSQL are been replaced with NoSQL like MongoDB, Cassandra, etc. Since the user access is from various gadgets like PC, tablets, mobile devices, interface should be neutral to any technology which creates bottleneck. The conventional OS and browsers will be replaced by Android, iOS for management of the user applications. Semantic web will be used increasingly for creation of meaningful content by relating document with the help of standards like RDF and RDFS (Schema) and ontology tools like OWL-Lite, OWL-DL, OWL-Full or DAML (DARPA Agent Markup Language) along with reasoning tools such as Jena, RacerPro, FaCT++, HermiT, Pellet etc. and modern query languages like SPARQL, DL Query, nRQL, RDQL, OWL-QL, SWARL etc. The modern libraries are envisaged as "Intelligent libraries" by introducing the trends and technologies discussed in this article.

References

1. FUJITSU, "FUJITSU Image Scanner fi-7280," 1995. [Online]. Available: http://www.fujitsu.com/global/products/computing/peripheral/scanners/fi/departmental/fi7280/. [Accessed 2015].
2. M. Obitko, "Introduction to Ontologies and Semantic Web," 2007. [Online]. Available: http://obitko.com/tutorials/ontologies-semantic-web/introduction.html.
3. W. S. Architecture, "W3C Working Group Note," 11 Feburary 2004. [Online]. Available: http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/#relwwwrest.
4. N. O. S. database, "Not Only SQL database," 2008. [Online]. Available: http://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL.
5. D. Space, "D SPACE," 3 Nonvember 2011. [Online]. Available: https://wiki.duraspace.org/display/DSDOC3x/Architecture.
6. U. Tokyo, "University of Tokyo Ishikawa Watanabe Laboratory," 2008. [Online]. Available: http://www.k2.t.u-tokyo.ac.jp/vision/BFS-Auto/.

SOUL 2.0

STATE-OF-THE-ART INTEGRATED LIBRARY MANAGEMENT SOFTWARE

http://www.inflibnet.ac.in/soul