Design and Fabrication of Information Architecture of a Digital Library

Prof. Anuradha Lyer, Ms. Deepali More, Mr. Sangameshwar and others

The paper would enable a technology-savvy reader to understand the technology of a digital library including service integration process, publishing process, data accessing process, wire-less network management process etc. along with the types and uses of architectural designs and integration activities.

- Prof. Shrikant Bhagali

Abstract:

This article is based on some secondary data from internet resources and books that enable digitization of Library, which is attained in symbiosis Institute of Management studies, A constituent of Symbiosis International University. This is basically a conceptual paper that speaks about various technologies, processes and architectures and functions that govern a digitized library system and how they are integrated to deliver services to the users. In SIMS, Symbiosis Institute of Management Studies, A constituent of Symbiosis International University, Library is getting digitized. The experiences and the exploration on the level of integrating technology and Architectures with processes and Functions in the system

Introduction:

Digital Libraries are designed to access the myriad of forms of knowledge from anywhere at anytime effectively and efficiently. For a user-friendly environment, we need a common infrastructure which is highly adaptive, and scalable. Invariably such architecture, will demand, state of the art technology and tools with innovative applications of such models and frameworks for a customizable and feasible solution for instant access to digital data. Seamless integration of the ideas technologies and innovative application models will enable the construction of coherent modular digital library infrastructure. The paper discusses the architectural approaches, processes, technologies, tools and activities that integrate to form a digital library.

The structure of information in a digital library

The information in a digital library is to be organized effectively to solve the queries. In a digital library, information is stored as a digitized map, a section of text, a Web page, a scanned photograph, etc. In digital form, each basic unit is a sequence of bits of data. However, users refer to material at a higher level of abstraction. Common English terms are used to refer to different types of items. They may have different formats, different usage restrictions; in spite of which users consider them equivalent.

Structuring Information has a few Issues.

Digital documents and images are frequently related to other materials by **relationships, such as** - a digitized text may consist of pages, chapters, front matter, and an index, illustrations. A computer program can be assembled from many files, with complex rules of inclusion. In the World Wide Web, a typical item may include several pages of text, with embedded images, and links.

The same item may be stored in several digital **formats.** Mostly, these formats are exactly equivalent and it is possible to convert from one to the other such as, an uncompressed image and the same image stored with a loss-less compression. At other times, the different formats contain different information, such as, different notations of a page of text in SGML and PostScript formats.

Different **versions** are created continually as these digital objects are easy to change. Versions differ by a single bit or many bits. A scanned photograph may have a high-resolution archival version, a medium quality version, and a thumbnail.

Every element of digital information will have different **rights and permissions.** The user accesses the material depending on the characteristics of **computer systems and networks** and the size of the document. Also, type of work differs from user to user who connects from a high-speed network over a dial-up line.

There is innate need for information architecture for information storage and management that specific codes can understand and execute the query from the user.

The information architecture must enable the regular activities automated and permit the staff to focus on decision-making aspects practically. Library Application software must have flexibility. This flexibility is incorporated into the information management systems from the types of queries, types of data being accessed, kinds of users, their network type and the bandwidth availability and the sequence of queries and the network modes that execute these queries practically.

The information architecture is developed based on economic, social, and legal frameworks of the system in which it is being established. The

architecture takes into account, the information being shared across unsecured networks and cross national boundaries for an internationally recognized university. Enterprise applications used under such a technical framework support such requirements.

Data types in Information Architecture

The information architecture is primarily based on data types, structural metadata, and meta-objects. A **data type** explains technical properties of data, such as format, or method of processing. **Structural metadata** is a data of the data that describes the types, versions, relationships and other characteristics of digital materials. A **meta-object** is an object that provides references to a set of digital objects.

Each item of data has an associated data type. The type specifies that the data have a certain format, such as JPEG, HTML, and ASP etc. All metadata are encoded. No semantic information is included in any name that is not encoded separately as metadata. Pointers are given to individual items of intellectual property. Whenever an item of information might be used on its own, it is given its own handle and made into a separate digital object. So that, an item may be accessed independently. This provides maximum long-term control and flexibility. If a digitized text contains illustrations that could potentially be used independently, each illustration is made into a separate digital object with its own handle.

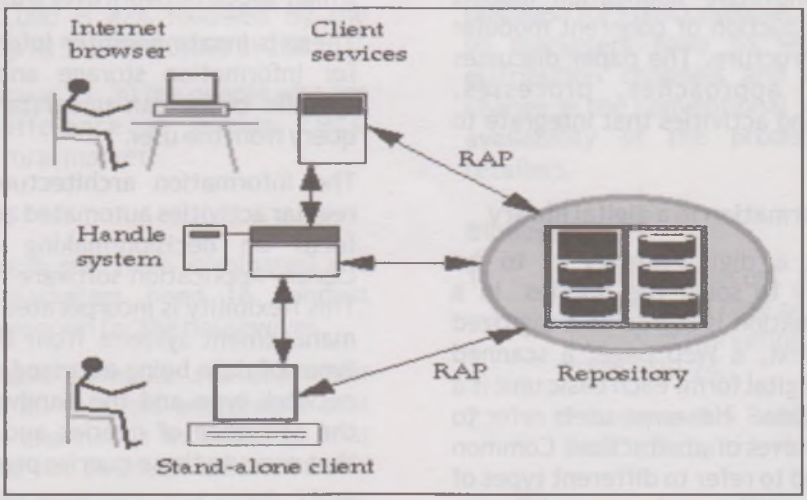
Meta-objects are used to aggregate digital objects. Complete metadata about a single piece of information may exist in several places within a repository, in external catalogs, indexes, or in searching aids. Maintaining links to all the metadata is a huge task. Handles are used to identify items listed in meta-objects. A meta-object contains a list. This enhances a robust, flexible structure that allows subsequent reorganization of the collection with minimal effort.

Interpretation of these rules is done with a trade-off between a powerful representation of information, which is flexible and laborious to manage. The architects of the information structure cannot dictate such decisions. The system provides straightforward methods for curators to decide how best to manage collections. The work concentrates on digital library entities that are converted from physical formats, such as photographs and printed articles. The architecture is designed to be generic. Digital objects store static or dynamic information; they can be archived for perpetuity or have a transitory life. Access to a digital object in a repository will normally require the execution of a program of arbitrary complexity. Repositories, themselves, may be within mobile agents. In our future work, we aim to extend the richness and variety of information in the digital library architecture by continuing to build upon the simple building blocks of digital objects, handles, and repositories.

Interfaces: Repository Clients

Repository clients here are thin client configuration with Linux server network system, used to locate and manipulate digital objects contained in repositories. The figure below describes the relationships between the repository, the repository clients, and the handle system.

This above system communicates directly with the handle system using the handle client library. Indirect user interfaces are now being used exclusively. They consist of an Internet browser for interactions with the user and client services. For the pilot, the client services are sets of cgi-bin scripts.



User interface architecture - The figure shows the two different types of clients that were implemented for the pilot repository.

User Interfaces for Librarians and System Administrators

For the pilot system, a user interface that has the features required. The user interface is implemented as a set of client services in this case. These services are of the interface has the following features.

- Runs on any computer that can support Netscape Navigator 2.0. or internet explorer or Firefox browser
- Supports the entire RAP methods implemented for the pilot repository.
- Provides an administration interface to create, edit, view, and delete digital objects.
- Provides a repository selection mechanism.
- Resolves digital object handles using the handle client library.
- Maintains a session with the repository.
- Easily configurable to support browser helper applications for unsupported data types.
- Provides for flexible viewing of metadata.
- Performs basic metadata verification for new objects that are deposited.
- Provides a mechanism for selecting which data elements to access for viewing and editing.

The limitation is the connection between the browser and the client services. The browser adds to overhead and potential security problems. Interface is not compatible as a stand - alone interface. These are temporary problems and should all be overcome in the prototype system. This structure of information architecture is very good for providing access to sets of digital objects that are in a true hierarchy. However, flexibility is less. All links are equal and arbitrarily complex structures can be created.

Processes of a Digital Library

Service Integration process: Digital libraries primarily function on autonomy and distribution of information providers data servers - at different nodes. Accessing information deliberates the integration of services into mega-applications, resulting in electronic workflow processes that manage and control consistency and accuracy of data search and data flow systems with self-configuration and flexibility, at application as well systems level.

Publishing Process: Services of a digital library is sought through service repositories to public. Subscription techniques make the information within digital libraries be available through replicated information sources. These technologies change with time and digital libraries have to update themselves on such techniques, software and schemata, every now and then to enable problem free quicker service to public. Usually the infrastructure provides the suitable mechanisms to distribute and co-ordinate updates to technology and tools in a kinetic environment.

Data Accessing Process: The efficiency of accessing information accelerates within a digital library, when it's replicated at several nodes in the network. Duplicate copies are made available at different data servers due to independent upload of data at different nodes. Whenever the data gets changed at one node, it reflects and updates other nodes as well simultaneously. The cost benefit analysis is usually done, to assess the trade off between update costs and freshness of data. Sophisticated mechanisms to trade, in an application-specific way, update costs for the freshness of data need to be provided by the digital library infrastructure.

Wireless Network Management Processes: Devices such as laptops, personal digital assistants, handheld PCs, and smart phones, enable a wide range of new digital library applications. These devices stress on the management of network through LAN as well Wi-Fi network technologies. This combination makes the digital Library network pervasive and permeable. The interfacing Middleware in the architecture supports flexibility and quickness of data access. User management, Profiling, security administration and proxy management get embedded into the architecture. Wireless devices enable the usage of sophisticated visualization techniques and processes to present digital publications adequately on limited displays.

The Digital Data flow Management Architectures and the above listed key processes need one of the following at least, for achieving best results for the users. They are,

Architectural designs:

There are three types of architectures.

1. **P2P Architectures:** Digital Data and documents of a digital library of an information provider are autonomous and they never get integrated under single source. There are mechanisms to retain this autonomy as well couple them loosely amongst themselves. This enhances collaborative data sharing among the network nodes (e.g., for annotations and recommendations about DL contents). Peer-to-peer (P2P) architectures permit such loosely coupled integration. Different aspects of peer-to-peer systems (e.g. indexes, and P2P application platforms)

are integrated to provide an optimal infrastructure for digital libraries.

2. **Grid Architectures:** Some specific services such as, calculation of certain features of multimedia documents to support content-based search within digital libraries are complex and computationally intensive. Grid computing architectures allow for sophisticated load balancing between nodes and work on a framework with a cluster of components. The service grid incorporates the handling and control of shared resources, well integrated into an infrastructure for digital libraries.

3. **SO Architectures:** There are SOA mechanisms that provide services based on the semantics and usage of the digital data and documents. Typically like that of an SOA platform of an ERP package. In the case of web services, descriptions of the specified services using service description languages are stored in service registries. These elements get integrated as building blocks into a digital library. Common service interfaces are defined and constructed based on existing service standards of a digital library to facilitate composition of entire service spectrum.

Functions of Architectures:

These services that make Digital library data and documents become available to tailor the type of data and documents and implement, for instance, appropriate index structures.

XML Storage: Access methods for digital documents in XML stores will provide the basis for mediation within P2P information architectures. XML is an efficient and effective annotation language, that emerge today. IR techniques over XML sources, shredding and selective indexing for fast retrieval, clustering, replication management, and transformation aimed at improved transport and platform specific delivery require more attention, in such an architecture.

Multimedia Access: The most important service of a digital library is the maintenance and retrieval of documents of various types under different search scenarios. Due to the distributed nature, services from different nodes must be able to interact with each other. To this end, well-designed service interfaces are required to ease integration of different providers featuring, extraction algorithms, indexing services, and retrieval engines.

Digital Rights Management: Legal aspects of publications on digital Library and the security issues have to be addressed efficiently. Digital rights, IPR have to be enforced. This involves support for business models such as pay-per-view or subscriptions. Peer-to-Peer (P2P) architectures are not so good in this regard from their application in the music industry context, and this situation is getting overcome to facilitate business development. Replication is a process that should be legalized.

Security and Certification: Peer to peer architecture requires authentication and authorization of users, authors, content providers, and reviewers as well as digital signatures for documents, to ensure the consistency, quality, and reliability of digital libraries. The results of various activities of the architecture cluster get integrated and evaluated in concrete settings, demonstrator systems and building blocks get combined. All the solutions will be developed in a sample application domain by clusters, to evaluate different digital library platforms and infrastructures.

Medical Information Systems: E-Health Warehouses are also a type of a digital Library. Information about patients is made available from different distributed and autonomous information providers. This information is integrated at application level in order to provide electronic patient records. Appropriate service interfaces of building blocks, workflow management system for application development, P2P infrastructures, replication and freshness, security and certification are integrated to achieve the goals of the digital library.

Digital Library Architecture Integration Activities:

Such an integration of activities and processes in a Digital Library Architectures cluster will address network and basic services architectures that allow integrated access to distributed digital libraries. This integration addresses, Development of surveys that collect the most significant contributions and promises in Digital Libraries Architectures; Developments of prototype software modules and components for web services, multiple service composition and management, wireless connectivity and Test of the solutions on a prototype ongoing application.

The cluster to enable the achievement of the above goals will support the following research activities.

Surveying the State of the Art:

Fundamental architectural problems in digital libraries demand adoption of new networking architectures, new standards, and integration of system components into library application software and workflow process, and integration with transmission media. The cluster surveys subjects, identifies solutions, technologies and promising scientific results.

Cluster brings in, Replication and Freshness of Data a mandatory mechanism for digital libraries, Security and Certification of digital documents, Network Architecture management. It gets in Collection level descriptions to enable information discovery and assists with information management within service registries.

P2P architecture enables loosely coupled integration of different facets of an application. Different aspects P2P systems such as indexes, application platforms, etc get combined and integrated into an infrastructure for digital libraries. Wherein, grid computing architectures allow for enhanced load balancing strategies within a cluster of components. Service grid controls shared resources, and integration of infrastructure for digital libraries. The cluster studies and manages common protocols for generic service models that provide appropriate descriptions of the available services in a digitized library set up. Web services get integrated as building blocks into digital libraries to provide access to individual services and to define common services. Applications that utilize digital libraries of autonomous information providers or applications that manage and control the consistency of a digital library, existing services of a digital library get integrated into all kinds of workflow processes. At application as well at systems level, different aspects of workflow management systems such as self-configuration and flexibility, high availability and scalability, evolve. The cluster in a distributed digital library application, in fact develops prototype of a system that demonstrates the feasibility of solutions offered.

Mobile Information Components:

The cluster fabricates a demonstrator that embeds innovative solutions to the access component of digital libraries using a combination of wired and wireless connectivity. The demonstrator incorporates a specific middleware that adapts the content to the limitations of all handheld devices.

Conclusion:

Design and Development of a Digital Library System, deliberates the understanding of technological components, their applications, architectures, processes, workflows, and the services, end users, the security and legality of the digital documents that are being accessed day and night. The front end, back end, and the processes and their complete integration without any blocks anywhere in the system make a digitization of a library project effective and successful.

References:

1] Computer Science Technical Reports Project (CSTR), 1992-96. (http://www.cnri.reston.va.us/home/cstr.html)

[2] Robert Kahn and Robert Wilensky, "A Framework for Distributed Digital Object Services", May 1995. (http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html)

[3] Key Concepts in the Architecture of the Digital Library by William Y. Arms, D-Lib Magazine, July 1995(http://www.dlib.org/dlib/July95/07arms.html)

[4] Carl Lagoze and David Ely, "Implementation Issues in an Open Architecture Framework for Digital Object Services", 1995. Cornell Computer Science Technical Report TR95-1540.

[5] Carl Lagoze, Robert McGrath, Ed Overly, Nancy Yeager, "A Design for Inter-Operable Secure Object Stores (ISOS)", 1995. Cornell Computer Science Technical Report TR95-1558.

[6] CNRI, "Handle Home Page", 1995-97. (http://www.handle.net/)

[7] Caroline R. Arms, "Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress". D-Lib Magazine. Part 1, April 1996 (http://www.dlib.org/dlib/april96/loc/04c-arms.html), Part 2, May 1996 (http://www.dlib.org/dlib/may96/loc/05c-arms.html).

[8] Library of Congress, American Memory Background Papers and Technical Information, 19595-96. (http://lcweb2.loc.gov/ammem/ftpfiles.html)

[9] Carl Lagoze, "The Warwick Framework a Container Architecture for Diverse Sets of Metadata". D-Lib Magazine, July/August 1996. (http://www.dlib.org/dlib/july96/lagoze/07lagoze.html)

[10] Department of Computer Science, Stanford University, "Stanford Digital Library Testbed Development". (http://www-db.stanford.edu/~testbed/)

[11] Network Of excellence On Digital Libraries, "Digital Library Architecture". http://dbis.cs.unibas.ch/delos_website/