United States Educational Foundation in India,
DRTC/Indian Statistical Institute,
DLIS/University of Mysore
Joint Workshop on Digital Libraries
12th – 16th March, 2001

Paper: G

Metadata Architecture for Digital Libraries: Conceptual framework for Indian Digital Libraries

Madhusudana Rao CR

Centre for Development of Advanced Computing,

Bangalore

Email: *madhu@cdacb.ernet.in*

Abstract

This paper describes approach of development of Metadata solution for digital library architecture for resource description and retrieval. This deals with the concept of Metadata [2], the different Metadata standards (Dublin core in particular [5]), Digital library environment, computer network capabilities etc. This paper also discusses two of the Digital Library architecture protocols, for resource description and retrieval. They are STARTS (Stanford Protocol Proposal for Internet Retrieval and Search) [8] and SODA (Smart Objects and Dump Archives)[13] architecture to arrive at a possible protocol that would help to build Indian Digital Libraries [5]. While proposing the new architecture the existing Indian environment with respect to information sources and user's query of the information sources [5.1], which are feasible for launch of this protocol for information processing and retrieval has been dealt with. This is a pilot study which the author has done while doing his Fulbright fellowship in the College of Library Information Studies, University of Maryland, College Park, MD during 1999-2000.

1 Introduction

As of now, there are a large number of superb digital libraries existing in the world, all of which are, unfortunately, vertically integrated and presenting a monolithic interface to their users [13]. In other words, if any user wants to make a search on any given topic, he has to go through each digital library search interface to retrieve what he wants. A user expects to locate his resource from a variety of digital libraries using only one interface. Even for a Digital Library Service Provider to build a successful digital library, the above interface would be an ideal one for a large community of users. For that, there have to be standard methods like universal protocols to interact with archives and digital objects. While considering user search and retrieval of relevant documents, they are available everywhere both vertically i.e. internal networks, and horizontally across the Internet. The source contents are often hidden behind search interfaces. As said in the beginning these interfaces vary from source to source. Also, the algorithms with which the associated search engines rank the documents in a query result are often incompatible across sources[8]. This paper discusses the issues related to defining a protocol for networked environment to choose the best sources to evaluate a query, evaluate the query at these sources and retrieve the best possible query results while merging different sources.

To achieve the above goal one needs to understand the following things:

1.1 For Information sources one needs to take into account (13)

- Identifying the user group
- Identifying the information sources
- Negotiating with different information sources
- Creating Indices by resource description format i.e. Metadata

1.2 For search and retrieval (8)

- Choose the best information source to evaluate the query
- Evaluate the query at these sources
- Merge the query results from these sources

There are several approaches that exist to achieve the above objective to retrieve a source for a given query. There are a number of reference implementation protocols that exist for each of these issues which have been identified above. In this paper, of the existing protocols, two have been identified and studied, which have been reported earlier and found feasible to adopt in this study for arriving at a new protocol for the Indian digital library environment.

To understand these protocols one needs to introduce a few concepts like User, User Query, Information Source, Networks, Internet, Interface, Search, Retrieval, Metadata etc. In order to focus on the critical issues, this paper details the concept of Metadata and two protocols that have been identified to arrive at a new protocol. For practical purpose it has been assumed that the readers are aware of other key concepts, which have been mentioned above.

2 Metadata

2.1 Introduction

Metadata describes an information resource. The term "meta" comes from a Greek word that denotes something of a higher or more fundamental nature (11). The simplest useful definition of metadata is "structured data about data."(5) This very general definition includes an almost limitless spectrum of possibilities ranging from human-generated textual description of a resource to machine-generated data that may be useful only to software applications.

The term metadata has been used only in the past 15 years, and has become particularly common with the popularity of the World Wide Web. But the underlying concepts have been in use for as long as collections of information have been organized. Library catalogs represent a well-established variety of metadata that has served for decades as collection management and resource discovery tools (5).

In general all information objects, regardless of the physical or intellectual form they take, have three features - content, context and structure - all of which can be reflected through Metadata. Content relates to what the object contains or is about, and is intrinsic to an information object. Context indicates who, what, why, where, how aspects associated with the object's creation and is extrinsic to an information object and Structure relates to the formal set of associations within or among individual information objects (7).

Library metadata development has been first and foremost about providing intellectual and physical access to content. Library metadata includes, indexes, abstracts and catalog records created according to cataloguing rules and structural and content standards such as MARC, AACR, LCSH, etc.(7)

To give an example, if you've ever completed a large and difficult jigsaw puzzle, you'll be familiar with that particular moment of grateful revelation when you find that two sections you've been working on separately actually fit together. The overall picture becomes coherent, and the task at last seems achievable.

Something like this seems to be happening in the puzzle of "content metadata". Two communities - rights owners on the one hand, libraries and cataloguers on the other -- are staring at their unfolding data models and systems, knowing that somehow together they make up a whole picture. Metadata are another level of content to librarians, but a means to the content for users. Not only do digital librarians face challenges in standardizing metadata to insure interoperability across digital libraries, but the range and distinctiveness of metadata are problematic. In some cases, it is only the metadata that is made available digitally. In such cases, users search through pointers and must acquire the primary information physically or through a different (e.g., fee-based) system. Such libraries are more properly considered as referral services rather than digital libraries. In more typical cases, metadata for objects of different granularity (e.g., titles for collections and titles for single objects) are mixed together on computer displays with full text or objects. In physical libraries, the card catalog or OPAC is physically distinct from the items on shelves. These distinctions are difficult to make in electronic environments because everything is displayed on the same physical screen; thus the boundaries between metadata and primary data are often blurred (14).

The following facts are to be noted before understanding about Metadata:

- Metadata does not have to be digital
- Metadata relates to more than the description of an object
- Metadata can come from a variety of sources
- Metadata continue to accrue during the life of an information object or system
- One information object's metadata can simultaneously be another information object's data (7).

2.2 Importance of metadata

Metadata is used primarily as intermediate steps to retrieving content. Creating new types of surrogates for objects to allow users to quickly preview and browse content is the challenge. In this context the concept of Metadata will come to the rescue for retrieval aspect and has become a more familiar theory with the advent of Internet and WWW. In view of the huge size and explosive rate of growth of the WWW, it is clear that catalogs of some kind would be invaluable in helping users discover relevant information resources. Unfortunately, neither the Internet nor the WWW were

originally designed with cataloging of their contents (6). But as the days passed, tools were designed to address the resource location problem and help to make sense of Internet's vast information resources (15). They are directories of listing of network resources and search engines.

The development of the WWW and other networked digital information systems has provided information professionals with many opportunities, while at the same time requiring them to confront issues that they have not had occasion to explore previously. Judiciously crafted metadata element sets, wherever possible conforming to national and international standards, have become the tools that information professionals are using to exploit some of these opportunities, as well as to address of the new issues (7). They are:

- ***Increased accessibility:*** Effectiveness of searching can be significantly enhanced through the existence of rich, consistent metadata. Metadata can also make it possible to search across multiple collections or to create virtual collections from materials that are distributed across several repositories.
- ***Retention of context:*** Repositories like Libraries, Archival, Museums do not simply hold objects. They maintain collections of objects that have complex inter relationships among each other and associations with people, places, movements and events. In the digital world it is not difficult for a single object from a collection to be digitized and then to become separated from both its own cataloging information and its relationship to the other objects in the same collection. Metadata plays a critical role in documenting and maintaining those relationships, as well as in indicating the authenticity, structural and procedural integrity, and degree of completeness of information objects.
- ***Expanding use:*** Digital information systems have a vital role of disseminating digital versions in very unique way beyond the barriers of geography and economics. It also can facilitate an almost infinite number of ways to search for information, present results, and even manipulate information objects without compromising the integrity of those information objects.
- ***Multi versioning:*** The existence of information and cultural objects in digital form has heightened interest in the ability to create multiple and variant versions of those objects. This process may be as simple as creating a high-resolution copy for preservation or scholarly research purposes and a low-resolution thumbnail image that can be rapidly transferred over a network for quick reference purposes.
- ***Legal issues:*** Metadata allows repositories to track the many layers of rights and reproduction information that exist for information objects and their multiple versions. Metadata also documents other legal or donor requirements that have been imposed on objects, e.g. privacy concerns or proprietary interests.
- ***Preservation:*** If digital information objects that are currently being created are to have a chance of surviving migrations through successive generations of computer hardware and software, or removal to entirely new delivery systems, they will need to have metadata that enables them to exist independently of the

system that is currently being used to store and retrieve them. Technical, descriptive, and preservation metadata that documents how a digital information object was created and maintained, how it behaves, and how it relates to other information objects will all be essential. It should be noted that for the information objects to remain accessible and intelligible over time, it will also be essential to preserve and migrate this metadata.

- ***System improvement and economics:*** Benchmarking technical data, much of which can be collected automatically by a computer, is necessary to evaluate and refine systems in order to make them more effective and efficient from a technical and economic standpoint. The data can also be used in planning for new systems.

Metadata is like interest -- it accrues over time (7). But the resources and intellectual and technical design issues involved in metadata development and management are far from trivial. Some of the key questions that must be resolved for sharing these resources are identifying which metadata schema or schemas should be applied in order to best meet the needs of the information creator, repository and users. For that one needs to know the metadata standards.

2.3 Standards for metadata on the web

In order for metadata to be as useful and cost-effective as possible, it is essential that its structure, semantics and syntax conform to widely supported standards, so that it is effective for the widest possible user community (6). There is no single international standard for metadata[1]. Several metadata schemes for digital information objects have been proposed, with different levels of complexity and richness from relatively simple formats, such as Dublin core to more complicated and richer formats like TEI (Text Encoding and Interchange). Clearly, the information structure and content of Web metadata records should capture the essence of the web resources they describe and facilitate the various tasks for which the metadata was devised. If there is a solution to the problem of resource discovery on the Web, it must surely be based on a distributed metadata model[6]. There are necessary protocols available for creating distributed and shared meshes of resource discovery models such as Z39.50 etc. What is required now is the widespread adoption of standards for metadata structure, content and authentication that will allow secure interoperability on semantic level.

There are many metadata standards that have evolved over the years. To name few:

- Dublin Core (5)
- IAFA templates
- WWW semantic header
- URS (Uniform Resources Citation)
- OCLC InterCat project
- TEI (Text Encoding and Interchange)
- Search Engine Meta tags
- Resource Description Framework

- EAD (Encoding Archival Description)
- GILS (Government Information Locator Service)
- Federal Geographic Data Committee
- Categories for the Description of Works of Art
- Museum Educational Site Licensing Project
- Common Object Request Broker Architecture

For this project Dublin Core has been identified because of its simplicity and standardized description of resource description.

3 Dublin Core (5)

3.1 What is the Dublin Core?

The Dublin Core metadata standard is a simple yet effective element set for describing a wide range of networked resources. Dublin Core metadata is specifically intended to support resource discovery. The elements represent a broad, interdisciplinary consensus about the core set of elements that are likely to be widely useful to support resource discovery. The original workshop was held in Dublin, Ohio, hence the term "Dublin Core" has been named to that effect.

The Dublin Core has become an important part of the emerging infrastructure of the Internet. Many communities are eager to adopt a common core of semantics for resource description, and the Dublin Core has attracted broad ranging international and interdisciplinary support for this purpose.

The Dublin Core standard comprises of fifteen elements, the semantics of which have been established through consensus by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields of scholarship. Each of these 15 elements is optional and may be repeated. Each element also has a limited set of qualifiers, attributes that may be used to further refine (not extend) the meaning of the element. Thus Dublin core has been further categorized as Simple DC or Unqualified DC and Qualified DC.

"Simple Dublin Core" is a term often used to describe Dublin Core metadata that uses no qualifiers. That is, the elements are expressed using just the 15 elements from the Dublin Core Metadata Element Set without any further information about encoding schemes, enumerated lists of values, or other processing clues. The term "Unqualified Dublin Core" is synonymous with "Simple Dublin Core."

"Qualified Dublin Core" is a term applied to Dublin Core metadata that employs additional information to increase the specificity of the metadata by refining the meaning, by specifying encoding schemes or controlled vocabularies, or to indicate a metadata value is a compound, or structured value. For example, a date may be further identified as a particular variety of date (date last modified, date published, etc.) and

might be encoded according to a particular scheme that assures that it can be interpreted unambiguously. A subject term that is the value of the subject element might be specified as having been selected from a particular controlled vocabulary such as the Dewey Decimal Classification.

3.2 Dublin Core has the following characteristics as its goals:

- ***Simplicity of creation and maintenance*:** The Dublin Core element set has been kept as small and simple as possible to allow a non-specialist to create simple descriptive records for information resources easily and inexpensively, while providing for effective retrieval of those resources in the networked environment.
- ***Commonly understood semantics*:** Discovery of information across the vast commons of the Internet is hindered by differences in terminology and descriptive practices from one field of knowledge to the next. The Dublin Core can help the 'digital tourist' -- a non-specialist searcher -- find his or her way by supporting a common set of elements, the semantics of which are universally understood and supported. For example, scientists concerned with locating articles by a particular author, and art scholars interested in works by a particular artist, can agree on the importance of a "creator" element. Such convergence on a common, if slightly more generic, element set increases the visibility and accessibility of all resources, both within a given discipline and beyond.
- ***International scope*:** The Dublin Core Element Set was originally developed in English, but versions are being created in many other languages. As of November 1999, there were versions in over 20 languages, including Finnish, Norwegian, Thai, Japanese, French, Portuguese, German, Greek, Indonesian, and Spanish. The Working Group on Dublin Core in Multiple Languages is coordinating efforts to link these versions in a distributed registry using the Resource Description Framework technology being developed by the World Wide Web Consortium (W3C).
- Although the technical challenges of internationalization on the World Wide Web have not been directly addressed by the Dublin Core development community, the involvement of representatives from almost every continent has ensured that the development of the standard considers the multilingual and multicultural nature of the electronic information universe.
- ***Extensibility*:** While balancing the needs for simplicity in describing digital resources with the need for precise retrieval, Dublin Core developers have recognized the importance of providing a mechanism for extending the DC element set for additional resource discovery needs. It is expected that other communities of metadata experts will create and administer additional metadata sets. Metadata elements from these sets could be linked with Dublin Core metadata to meet the need for extensibility. This model allows different communities to use the DC elements for core descriptive information which will be usable across the Internet, while allowing domain specific additions which make sense within a more limited arena.

3.4 The core elements of Dublin Core

The following are the elements and are listed in the order they were developed:

- ***Content***
  Coverage, Description, Type, Relation, Source, Subject and Title

- ***Intellectual property***
  Contributor, Creator, Publisher and Rights

- ***Instantiation***
  Date, Format, Identifier and Language

3.4.1 Dublin Core elements description

1. Title: The name given to the resource, usually by the Creator or Publisher.

2. Author or Creator: The person or organization primarily responsible for creating the intellectual content of the resource. For example, authors in the case of written documents, artists, photographers, or illustrators in the case of visual resources.

3. Subject and Keywords: The topic of the resource. Typically, subject will be expressed as keywords or phrases that describe the subject or content of the resource. The use of controlled vocabularies and formal classification schemas is encouraged.

4. Description: A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.

5. Publisher: The entity responsible for making the resource available in its present form, such as a publishing house, a university department, or a corporate entity.

6. Other Contributor: A person or organization not specified in a Creator element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organization specified in a Creator element (for example, editor, transcriber, and illustrator).

7. Date: A date associated with the creation or availability of the resource. Recommended best practice is defined in a profile of ISO 8601 (http://www.w3.org/TR/NOTE-datetime) that includes (among others) dates of the forms YYYY and YYYY-MM-DD. In this scheme, the date 1994-11-05 corresponds to November 5, 1994.

8. Resource Type: The category of the resource, such as home page, novel, poem, working paper, technical report, essay, dictionary. For the sake of interoperability, Type should be selected from an enumerated list that is under development in the workshop series.

9. Format: The data format and, optionally, dimensions (e.g., size, duration) of the resource. The format is used to identify the software and possibly hardware that might be needed to display or operate the resource. For the sake of interoperability, the format should be selected from an enumerated list that is currently under development in the workshop series.

10. Resource Identifier A string or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented). Other globally unique identifiers, such as International Standard Book Numbers (ISBN) or other formal names would also be candidates for this element.

11. Source: Source Information about a second resource from which the present resource is derived. While it is generally recommended that elements contain information about the present resource only, this element may contain metadata for the second resource when it is considered important for discovery of the present resource.

12. Language: The language of the intellectual content of the resource. Recommended best practice is defined in RFC 1766 (*http://info.internet.isi.edu/in-notes/rfc/files/rfc1766.txt*)

13. Relation: An identifier of a second resource and its relationship to the present resource. This element is used to express linkages among related resources. For the sake of interoperability, relationships should be selected from an enumerated list that is currently under development in the workshop series.

14. Coverage: The spatial and/or temporal characteristics of the intellectual content of the resource. Spatial coverage refers to a physical region (e.g., celestial sector) using place names or coordinates (e.g., longitude and latitude). Temporal coverage refers to what the resource is about rather than when it was created or made available (the latter belonging in the Date element). Temporal coverage is typically specified using named time periods (e.g., Neolithic) or the same date/time format (*http://www.w3.org/TR/NOTE-datetime*) as recommended for the Date element.

15. Rights Management: A rights management statement, an identifier that links to a rights management statement, or an identifier that links to a service providing information about rights management for the resource.

There are now a number of large-scale deployments of Dublin Core metadata around the globe. The official Dublin Core Web site lists 15 in North America and Mexico, in Europe and 12 across Asia and Australia. Some of these initiatives are on a national scale, for example the Australian Government Locator Service and the CCTA Government Information Service in the UK (6).

Although significant progress in raising awareness and increasing deployment of the Dublin Core has been made over the last few years, there is still a long way to go before it can begin to deliver on its promise of better resource discovery on the Web[6].

4 Digital Library Architecture

The two established architectures that has been identified for this study are:

- **SODA (Smart Objects and Dump Archives)**
- **STARTS (Stanford Protocol Proposal for Internet Retrieval and Search)**

4.1 SODA (Smart Objects and Dump Archives)

In the days of the Internet and WWW, Digital Libraries are an important source of many information queries and research areas. However, access to these DL's is not as easy as users would like. Digital libraries are partitioned both by the discipline they serve (for example Computer Science, Aeronautics, Physics etc.) and by the format of their holdings (technical reports, video, software, etc.). There are two significant problems with current DLs. First, inter disciplinary research is difficult because the collective knowledge of each discipline is stored in incompatible DLs that are known only to the specialists in the subject. The second significant problem is that although scientific and technical information consists of manuscripts, software, data sets, etc., the manuscript receives the majority of attention, and the other components are often discarded. A recent NASA study found that customers desire to have the entire set of manuscripts, software, data etc. available at one place. With the increasing availability of all-digital storage and transmission, maintaining the tight integration of the original information collection is now possible. (NASA report) (17).

To build a successful Digital Library, there have to be standard methods to interact with archives and digital objects. This SODA model proposes such a standard. In this model self-contained, intelligent and aggregate Digital Library objects exist that are capable of enforcing their own terms and conditions, negotiating access, and displaying their contents. These are specialized class of digital objects called "buckets". Once the client makes use of this service he would get a location of a bucket, it is up to the bucket to interact with him. In this model archives are simply collections of buckets characterized by some management policy that controls the publishing. It is the archive-owning organization that negotiates with the service provider or interface for access to the archive's buckets.

Thus while building information source of DL one has to: (13)

- Identify a user group
- Identify archives holding buckets of interest and individual bucket owners
- Negotiate terms and conditions with publishing organizations (archive and individual bucket owners)
- Create indices of appropriate subsets by interacting with buckets for their metadata
- Create Digital Library services such as search and browse
- Create user interaction services such as authentication and billing.

4.1.1 The SODA model

This model composed of three strategies:

- Digital Library services: The "user" functionality and interface: searching, browsing, usage analysis, citation analysis, selective dissemination of information, etc.
- Archive - managed sets of digital objects: DLs can poll archives to learn of newly published digital objects, for example.
- Digital object- the stored and trafficked digital content. These can be simple files (e.g. PDF or PS files), or more sophisticated objects such as buckets (described below).

The strategy mentioned in the (Fig.1) is called Smart Objects, Dumb Archive (SODA) model. Much of the traditional functionality associated with archives (terms and conditions, content display etc.) has been "pushed down" into the objects, making the objects "smarter" and the archives "dumber". The model has been implemented in NCSTRL using Dienst protocol (13).

In this model four concepts are involved. One can make different combinations to understand how it would fare and also to realize the existing protocol reference implementation. For example (13):

- SOSA: Smart Objects Smart Archives. Ex: none known
- SODA: Smart Objects Dumb Archives. Ex: NCSTRL+
- DOSA: Dumb Objects Smart Archives. Ex: NCSTRL
- DODA: Dumb Objects Dumb Archives. Ex: Any anonymous FTP servers.

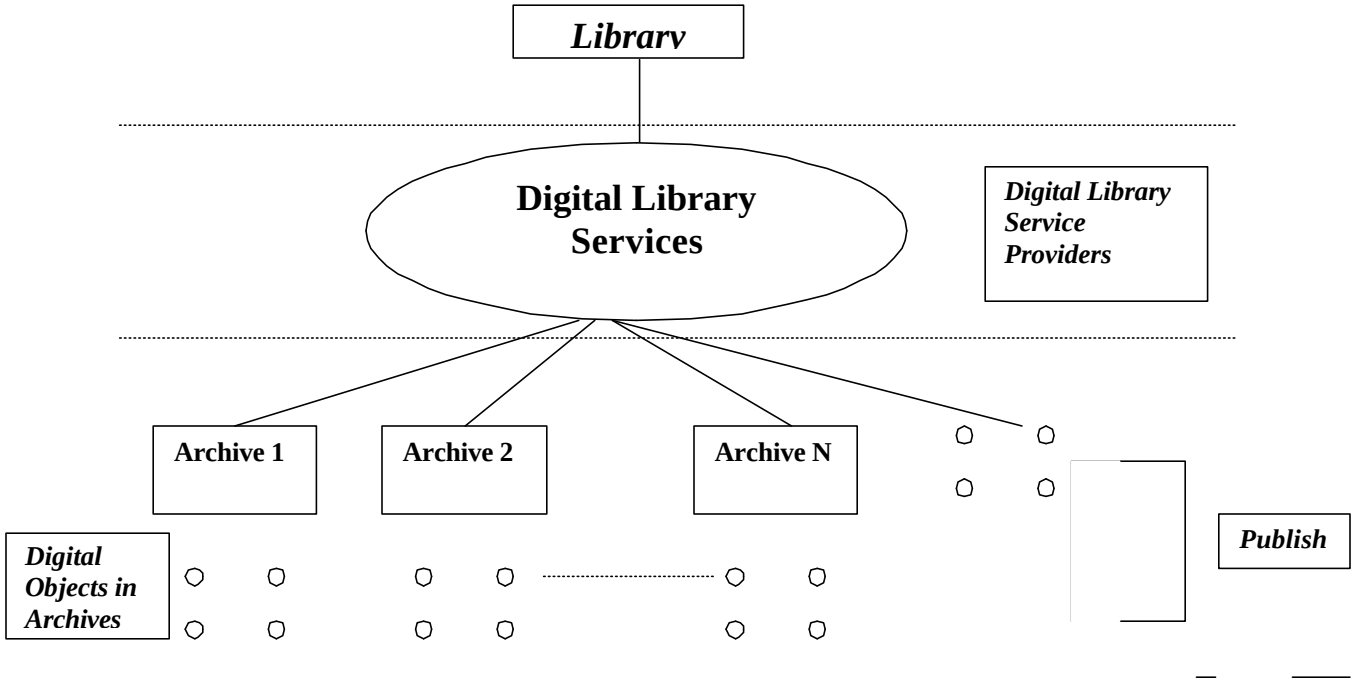

Fig. 1: The three DL's Strategy

4.1.2 Bucket architecture (17)

Buckets are object-oriented container constructs in which logically grouped items can be collected, stored, and transported as a single unit. For example, a typical research project at NASA Langley Research Center produces information tuples: raw data, reduced data, manuscripts, notes, software, images, video, etc. Normally, only the report part of this information tuple is officially published and tracked. The report might reference on-line resources, or even include a CD-ROM, but these items are likely to be lost or degrade over time. Some portions such as software can go into separate archives but this leaves researcher to re-integrate the information tuple by selecting pieces from multiple archives. Most often the software and other items, such as data sets are simply discarded. After 10 years, the manuscript is almost surely the only surviving artifact of the information tuple.

Large archives could have buckets with much different functionality. Not all bucket types or applications are known at this time. However, one can describe a generalized bucket as containing many formats for the same data item (PS, Word, PDF, Framemaker, etc.) but more importantly, it can also contain collections of related non-traditional materials (manuscripts, software, datasets, etc.). Thus, buckets allow the digital library to address the long standing problem of ignoring software and other supportive material in favor of archiving only the manuscript by providing a common mechanism to keep related products or sources together. A single bucket can have multiple packages. Packages can correspond to the semantics of the information (manuscript, software, etc) or can be more abstract entities such as the metadata for the entire bucket, bucket terms and conditions, pointers to other buckets or packages, etc. A single package can have several elements, which are typically different file formats of the same information, such as the manuscript package having both PostScript and PDF elements. Elements correspond to the syntax of a package.

All buckets have unique ids or handles, associated with them. Buckets are intended to be either standalone objects or to be placed in digital libraries. A standalone bucket can be accessed through normal WWW means without the aid of a repository. Buckets are intended to be useful even with repositories, which has no knowledge about buckets in general, or with the specific form of buckets. Buckets should not lose functionality when removed from their repository.

A high level list of bucket requirements include:

- A bucket is of arbitrary size
- It has a globally unique identifier
- It contains 0 or more components, called packages (no defined limit)
- A package contains 1 or more components called elements (no defined limit)
- An element can be a file or a pointer
- Both packages and elements can be other buckets (i.e. buckets can be nested)

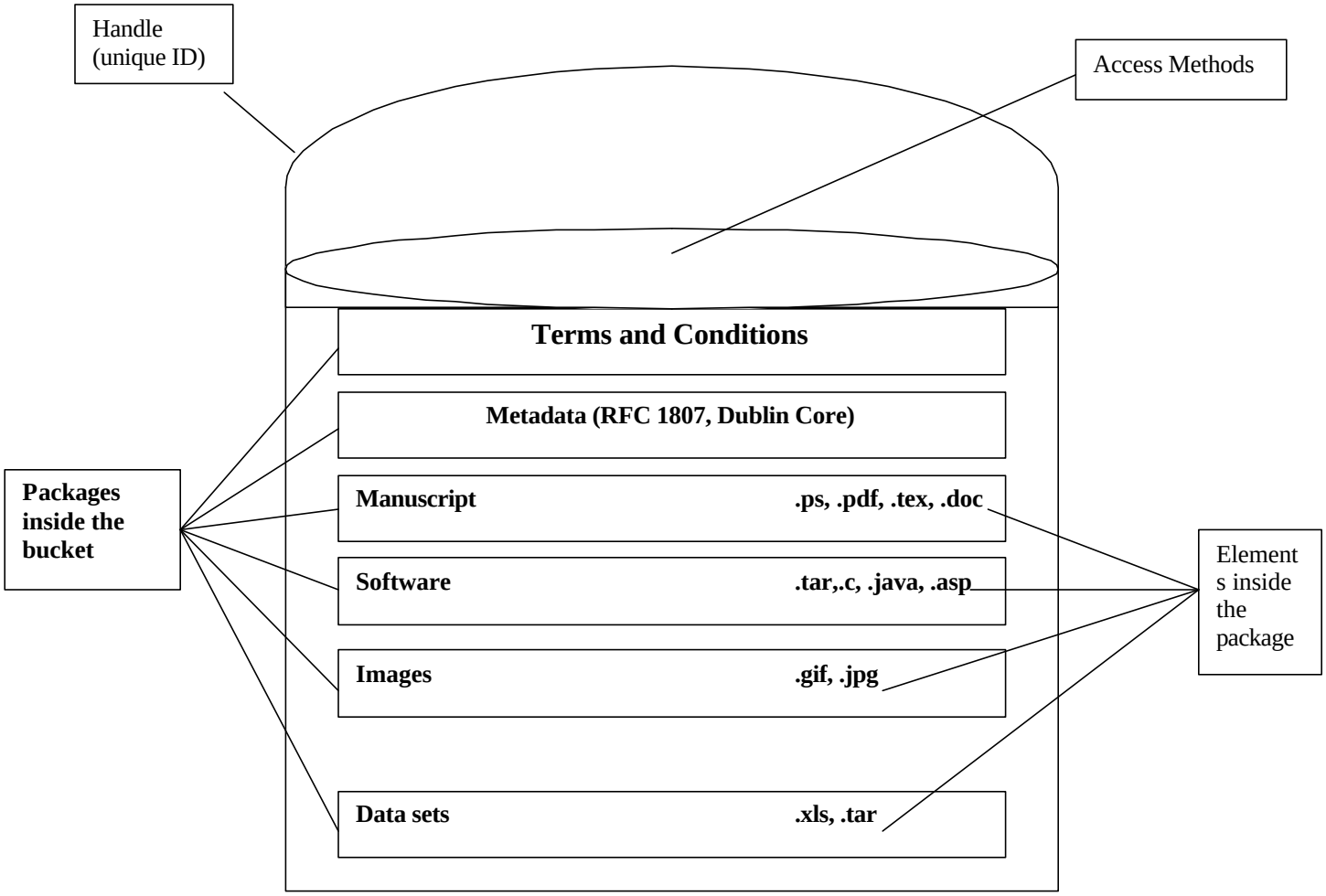

Fig. 2: Bucket Architecture

- A package can be a pointer to a remote bucket, package, or element (remote package or element access requires "going through" the remote hosting bucket)
- Packages and elements can be pointers to arbitrary network services, foreign keys to databases, etc.
- Buckets can keep internal logs of actions performed on them
- Interactions with packages or elements are made only through defined methods on a bucket
- Buckets can initiate actions; they do not have to wait to be acted on
- Buckets can exist inside or out of a repository

4.1.3 Bucket Tools (17)

A number of Buckets tools can be evolved based on the individual requirement and criteria. As of now NASA has brought out two tools for bucket use. One such tool is author tool, which allows the author to construct a bucket and another is Management tool, which provides an interface to allow site managers to configure the default settings for all authors at that site.

Bucket matching system is another tool where in archived objects (buckets) should handle as many tasks as possible. The protocol is such that communication mechanism exists for buckets to talk and exchange information with each other. The solution is such that if you have more than one bucket, each bucket would publish their metadata or some subset of it, in the Bucket Matching System (BMS). When a match or near match is found, buckets can either 1) automatically link to each other or more likely 2) bring the possible linkage to the attention of person, who will provide the final approval for the linkage.

4.1.4 Bucket Implementation

Old Dominion University and NASA Langley Research Center are developing NCSTRL+ to address the multi-discipline and multi-genre problems. NCSTRL+ is based on the Networked Computer Science Technical Report Library (NCSTRL) which is a highly successful digital library offering access to over 100 University departments and laboratories since 1994, and is implemented using the Dienst Protocol (17).

NCSTRL+ includes selected holdings from the NASA Technical Report Server (NTRS) and NCSTRL, providing clusters of collections along the dimension of disciplines such as aeronautics, space science, mathematics, computer science and physics as well as clusters along the dimension of publishing organization and genre such as project reports, journal articles, theses, etc (13).

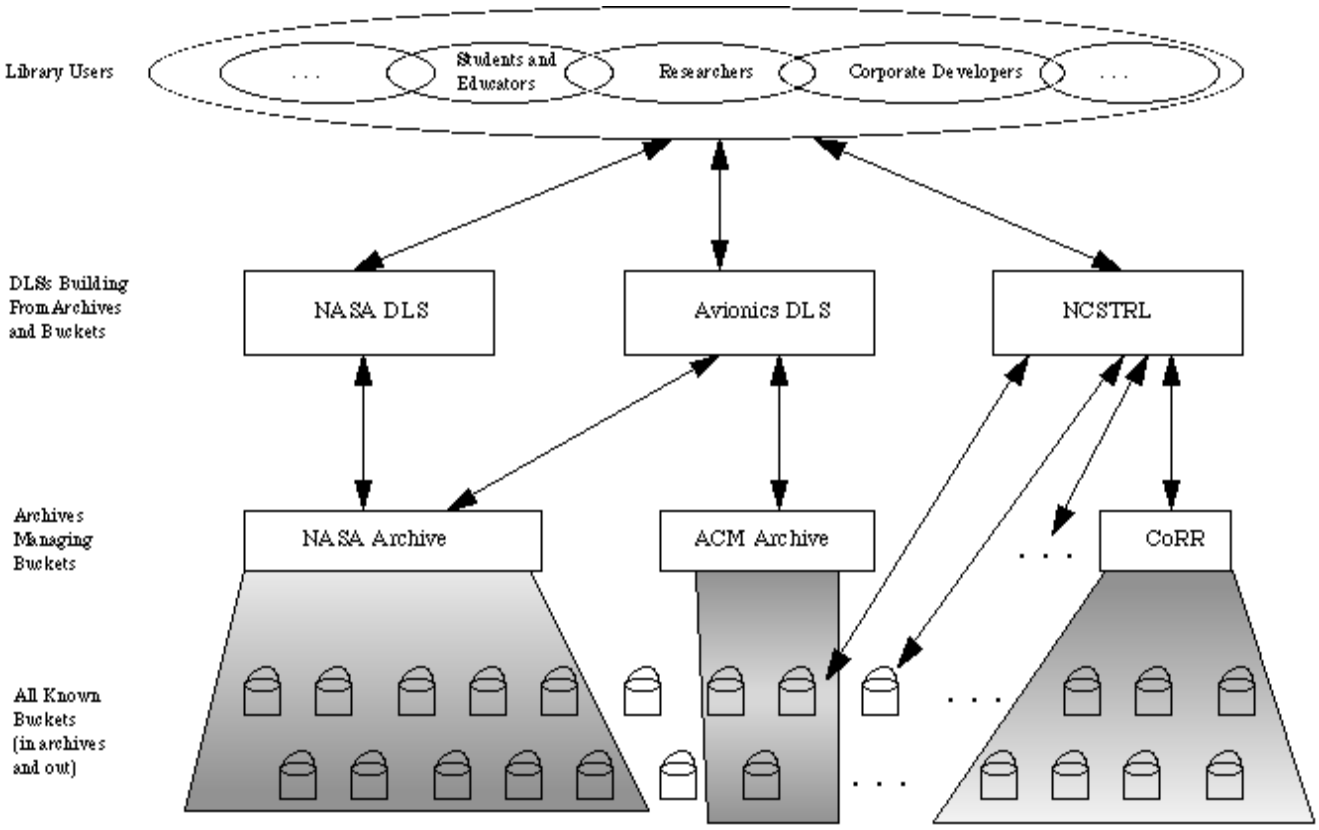

Fig. 3: SODA Publishing model

Bucket architecture provides a mechanism for logically grouping the various semantic data objects and various syntactic representations. The ability to keep all the data objects together with their relationships intact relieves the user from having to reintegrate the original information tuple from many separate archives. Buckets also provide a more convenient method for describing the output of research projects and a finer granularity for controlling terms and conditions within an archive. The aggregative aspects of buckets have already been implemented. The tools to make buckets easy to use and manage are being created.

4.2 STARTS (Stanford Protocol Proposal for Internet Retrieval and Search) (8)

Document sources are available everywhere, both within the internal networks of organizations and on the Internet. The source contents are often hidden behind search interfaces. These interfaces vary from source to source. Also, the algorithms with which the associated search engines rank the documents in query results are often incompatible across sources. Even individual organizations use search engines from different vendors to index their internal document collections. These organizations could benefit from unified query interfaces to multiple search engines, for example, that would give users the illusion of a single combined document source. Building

metasearchers (i.e., services that provide such a unified view of the multiple sources) is nowadays a hard task because different search engines are largely incompatible and do not allow for interoperability.

Given a query, a metasearcher has to perform (at least) three tasks to provide a unified interface over a (large) number of document sources:

- Choose the best sources to evaluate the query
- Evaluate the query at these sources
- Merge the query results from these sources

The existing search engines do not help with the three tasks above. In general, text search engines:

- Do not export information about the sources (the source-metadata problem)
- Use different query languages (the query-language problem)
- Rank documents in the query results using secret algorithms (the rank-merging problem)

To improve this situation, the Digital Library project at Stanford coordinated search engine vendors and other key players to informally design a protocol that would allow basic interactions of sources in the three areas above. This draft is based on feedback from people from Excite, Fulcrum, GILS, Harvest, Hewlett-Packard Laboratories, Infoseek, Microsoft Network, Netscape, PLS, Verity, and WAIS, among others.

In this architecture, there are (potentially large) numbers of resources. Each resource consists of one or more sources, and simply exports contact information for its sources. A source is a collection of flat documents (e.g., one may not consider any nesting of documents) with an associated search engine that accepts queries from clients and produces results. Sources may be "small" (e.g., the collection of papers written by some university professor) or "large" (e.g., the collection of WWW pages indexed by a crawler).

STARTS protocol is meant for machine-to-machine communication: users should not have to write queries using the proposed query language, for instance.

A metasearcher or any end client, in general, would typically issue queries to multiple sources. For this, a client will perform the following tasks:

- Extract the source list from the resources periodically (to find out what sources are available for querying)
- Extract metadata and content summaries from the sources periodically (to be able to decide, given a query, what sources are potentially useful for the query)

Given a user query:

- Issue the query to a source at a resource (Source 1 in the figure below), maybe specifying other sources at the resource where to also evaluate the query (Source 2 below)
- Issue the query to other promising resources
- Get the results from the multiple resources, merge them, and present them to the user

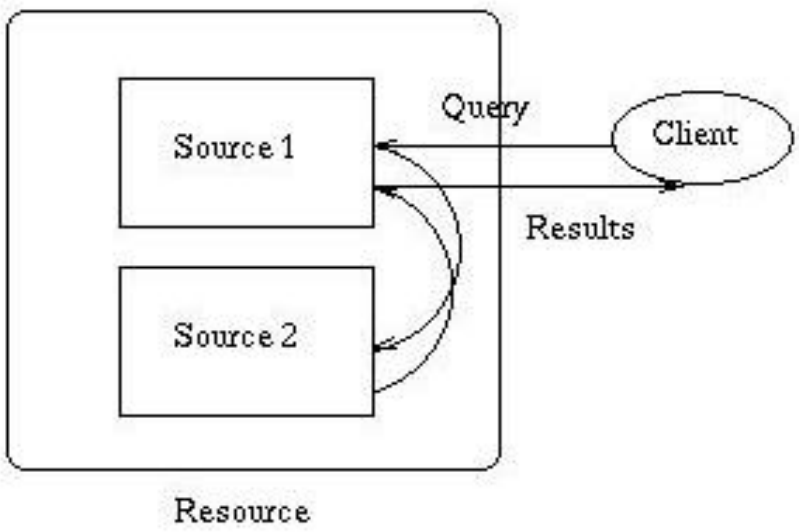

Fig. 4: STARTS Architecture

This protocol describes how to query sources and what information the sources export about themselves. It does not describe architecture for "metasearching," for example. However, it enumerates the facilities that a metasearcher would need from the sources in order to perform searches. Metasearchers often have to search across simple sources as well as across sophisticated sources. On the one hand, it is important to have some agreed-upon minimal functionality that is simple enough for all sources to comply. On the other hand, it is important to allow the more sophisticated sources to export their richer features. Therefore, our protocol keeps the requirements to a minimum, while it provides optional features that fancy sources can use if they wish.

The most relevant standards to this effort in terms of shared goals are the Z39.50 standard. Z39.50 provides most of the functionality that is feasible for this protocol. This proposal is much simpler, and keeping it simple was one of the main objectives of this effort. Other related efforts are Harvest and RDM. Both efforts provide a framework for querying and indexing multiple sources of documents. This effort is complementary in that it defines the pieces of information that sources should export to gatherers, and the query language and query-result format that the brokers should support, for example (using Harvest's terminology).

4.2.1 Query language/interface

The basic features of the query language that a source should support are based on a simple subset of the type-101 queries of the Z39.50-1995 standard.

A query consists of two parts:

- a filter expression, and
- a ranking expression.

A filter expression is Boolean in nature and defines the documents that qualify for the answer. The ranking expression associates a score with these documents and ranks them accordingly.

To achieve the above two objectives the following concepts have been well defined in the protocol

L-strings: An L-string is either a string or a string qualified with its associated language and, optionally, with its associated country. For example Garcia Molina (author), [en-US "behavior"] is an L-string meaning the string "behavior" represents a word in American English.

Atomic Terms: A term (e.g., (author)) is an L-string modified by an unordered list of attributes. An attribute is either a field or a modifier. For example, the term (date-last-modified > "1996-08-01") has field date-last-modified and modifier.

Complex filter expressions: This protocol will use operators to build complex filter expressions from the terms. If a source supports filter expressions, it must support all these operators: “and”, “or”, “and-not”, “prox” (proximity), specifying two terms, the required distance between them, and whether the order of the terms matters.

Example:

Consider two term's t1 and t2 and the following filter expression:

(t1 prox[3,T] t2)

The documents that match this filter expression contain t1 followed by t2 with at most three words in between them. "T" (for "true") indicates that the word order matters (i.e., that t1 has to be appear before t2).

Complex ranking expressions: The system uses operators to build complex ranking expressions from the terms. The "Basic-1"-type ranking expressions use the operators above ("and," "or," "and-not," and "prox") plus a new

operator, list, which simply groups together a set of terms. (The "Boolean" operators would most likely be interpreted as "fuzzy-logic"operators by the search engines in order to rank the documents.) If a source supports ranking expressions, it must support all these operators.

Each term in a ranking expression may have a weight (a number between 0 and 1) associated with it, indicating the relative importance of the term in the query.

Example:

The following ranking expression indicates that the term "distributed" is more important than the term "databases":

list(("distributed" 0.7) ("databases" 0.3))

Global settings: The system also defines the global settings such as stop words, default attribute set used in the query, default language used in the query sources etc.

4.2.2 Merging ranks

After receiving a query, the source reports the number of documents in the result. Also, since the source might modify the given query before processing it, the source reports the query that it actually processed.

Example:

Consider a source that does not support the ranking-expression part of the queries. Consider the query with filter expression:

((author "Garcia Molina") and (title "databases"))

and ranking expression:

list((body-of-text "distributed") (body-of-text "databases"))

If the source simply ignores the ranking expressions, the actual query that the source processes has filter expression:

((author "Garcia Molina") and (title "databases"))

and an empty ranking expression. This actual query is returned with the query results.

To merge the query results from multiple sources into a single, meaningful rank, a source should return the query result after considering the following issues:

- un normalized score,
- the id of the source(s) where the document appears,
- statistics about each query term in the ranking expression (as modified by the query fields, if possible) like term-frequency, term-weight, document-frequency, (this information is also provided as part of the metadata for the source), document size and document count.

4.2.3 Source metadata

To select the right sources for a query and to query them the system needs information about their contents and capabilities. This protocol proposes two pieces of metadata that every source is required to export: 1) a list of metadata attributes, describing properties of the source, and 2) a content summary of the source. Each piece is a separate object, to allow metasearchers to retrieve just the metadata that they need. Also, the protocol describes the information that a resource exports. This information identifies the metadata objects for the sources in the resource itself. For reference implementation the researchers have used Z39.50-1995 standard.

4.2.4 STARTS implementation

The reference implementation for STARTS is available in the Alexandria Digital Library (ADL) project. ADL is a research digital library project focussed on georeferenced/geospatial information and on geospatial data types such as maps, aerial photographs, remote sensing images, and data pertaining to particular geographical area.

ADL has two main collections for search and retrieval. They are ADL catalog and ADL Gazetter. The architecture allows ADL to accept multiple collections to represent a range of collection types with very different metadata schemas for each item description.

4.2.5 Conclusion

The brute-force method of searching the Internet by search engines such as AltaVista gathers and indexes "all" documents (i.e. information objects) on the network. STARTS on the other hand, first gathers metadata about collections and then selects a small set of collections to search directly. STARTS is more scalable than the brute-force method because it gathers collection-wide metadata rather than every individual document.

5 Indian digital library initiatives

5.1 Background

Few countries in the world have such an ancient and diverse culture as India's. Stretching back in an unbroken sweep over 5000 years, India's culture has been enriched by successive waves of migration, which were absorbed into the Indian way of life. It is this variety which is a special hallmark of India. Its physical, religious and racial variety is as immense as its linguistic diversity. Underneath this diversity lies the continuity of Indian civilization and social structure from the very earliest times until the present day. Modern India presents a picture of unity in diversity to which history provides no parallel. India is the largest democracy in the world, the seventh largest country in area, with a population of over 900 million people. Political process of election after election in which leader after leader has been replaced peacefully.

But it has a true history of democratic institutions: political parties come, new ones develop, old ones go away, a judiciary that works, a legislature that has real power, a press that is about as vibrant as any in the world. This is a country that is still heavily illiterate, but has voter turnout regularly of over 65 per cent, a remarkable history for any country, but particularly one that only got its independence 50 years ago.

India is an increasingly important part of the world economy. By almost any measure, it is already one of the world's largest economies. And it has one of the world's largest middle classes. The weakness of the Indian economy is its poverty.

India was the first and is still the largest exporter of Software among the developing countries (9). This has set a trend to get the best results of IT development in India. The feverish pitch which has started late 80's has never dwindled till date and now India is considered one of the major player of IT in the world. Although the growth of IT industry was slow and erratic in the beginning, exports particularly began to grow and now the industry has the capabilities of nurturing the best IT infrastructure with skilled manpower in a global environment.

In the world of Internet and World Wide Web, a lot of importance is given to heritage and religious contents. But offerings on Indian cultural aspects are not up to the mark, or are not comparable to those for western religions and even those available provide a distorted picture, as in the case of history. While western scholars have done excellent research on certain aspects of culture the soul is missing, but their work and method is widely available. For example, the *Vedas* happen to be the best and oldest available evidence of ancient Indian literary advancement and cultural heritage and their preservation is our foremost duty. Besides the word/sentence structure, the euphonic combination processes, the accent-related meaning variations, the diverse ways of word formations and usage, the etymological and exegetical aspects, recitation and tonal aspects etc. have profound influence on and can contribute to our knowledge and study of Indology, Linguistics (graphical, spoken and conceptual

forms), Life Sciences, Musicology, Medicine etc. Thus, preservation, study/research and propagation for posterity are the guiding objectives of any information sources.

As said earlier India has rich cultural heritage, components of heritage includes Indian art, Indian paintings, Indian sculpture, Indian religions, language etc.

The story of Indian art is also the story of the oldest and the most resilient culture on earth. It is seen as an amalgamation of indigenous and outside influences, yet having a unique character and distinctiveness of its own. Indian art is also an art of social, political and religious influences. It changed and evolved with the evolution of a civilization, which was full of remarkable innovations in all areas of artistic expression. Indian art features spirals and curvaceous lines, vines and tendrils, round-figured goddesses, circular amulets, colored gemstones, arches and domes, halloed deities, crescent moons, and the globe of the sun. Indian sculptures and paintings depict the diversity, colour and spontaneity of this country and are representations of the all-encompassing nature of Indian culture.

Indian paintings provide an aesthetic continuum that extends from the early civilization to the present day. This form of art in India is vivid and lively, refined and sophisticated and bold and vigorous at the same time. From being essentially religious in purpose in the beginning, Indian paintings have evolved over the years to become a fusion of various traditions, which influenced them.

The story of Indian art and sculpture dates back to the Indus valley civilization of the 2nd and 3rd millennium BC. Tiny terra-cotta seals discovered from the valley reveal carvings of peepal leaves, deities and animals. These elemental shapes of stones or seals were enshrined and worshipped by the people of the civilization. Two other objects that were excavated from the ruins of the Indus valley indicate the level of achievement that Indian art had attained in those days. The bust of a priest in limestone and a bronze dancing girl show tremendous sophistication and artistry.

In India, religion is a way of life. It is an integral part of the entire Indian tradition. For the majority of Indians, religion permeates every aspect of life, from common-place daily chores to education and politics. Secular India is home to Hinduism, Islam, Christianity, Buddhism, Jainism, Sikhism and other innumerable religious traditions. Hinduism is the dominant faith, practised by over 80% of the population. Besides Hindus, Muslims are the most prominent religious group and are an integral part of Indian society. In fact India has the second largest population of Muslims in the world after Indonesia.

Common practices have crept into most religious faiths in India and all communities share many of the festivals that mark each year with music, dance and feasting. Each has its own pilgrimage sites, heroes, legends and even culinary specialties, mingling in a unique diversity that is the very pulse of society.

In order to preserve these rich aspects of culture, the Indian Government has initiated the National Information Infrastructure (NII) with the objective of collecting information about this rich heritage, preserve it and retrieve it for those who seek information about ancient Indian tradition. In order to harness the rich traditions with the tools of Information Technology, NII has adopted several programs to meet this objective. One such program is to create a Digital Library of Indian Heritage. Digital Library of Indian Heritage is an important component. Like this there are initiatives in other key areas such as Agriculture, Water Management and many initiatives in Science and Technology areas such as Telecommunications, Information Systems, etc.

For implementation purpose if you look at the infrastructure and tools to felicitate DL initiatives, although communication and computer network facilities have come of age, there is still a big gap between the have's and the have-not's. Contrary to this India is emerging as a super power in IT infrastructure next to the US. Harnessing this technology to make our culture sources for retrieval is not so easy. One of the major hurdles in this direction is language. It has been said that India is a living Tower of Babel! There are fifteen national languages recognized by the Indian constitution and these are spoken in over 1600 dialects. Add to this a population of over 900 million today, and the remark would seem to be true. India's official language is Hindi in the Devanagri script. However, English continues to be the official working language. For many educated Indians, English is virtually their first language, and for a great number of Indians who are multi-lingual, it will probably be the second.

The country has a wide variety of local languages and in many cases the State boundaries have been drawn on linguistic lines. Besides Hindi and English, the other popular languages are Assamese, Bengali, Gujarati, Kannada, Kashmiri, Konkani, Sanskrit, Sindhi, Tamil, Malayalam, Marathi, Punjabi, Oriya, Telugu and Urdu. Some Indian languages have evolved from the Indo-European group of languages and these were the languages of the Aryans who invaded India. This set is known as the Indic group of languages. The other set of languages are Dravidian and are native to South India, though a distinct influence of Sanskrit and Hindi is evident in these languages. Most of the Indian languages have their own script and are spoken in the respective states along with English.

5.2 Proposed DL architecture

While keeping in mind the availability of Networked information sources with loose connectivity bandwidth and multilingual data banks of sources, this paper proposes a hybrid architecture, which suits the present day needs and is also feasible to adopt in an environment where it needs to deploy. It proposes the combination of both SODA and STARTS architecture for implementation in Indian digital library environment.

For information source management it is proposed to take SODA model and for information management and retrieval aspects, the STARTS architecture is considered suitable.

5.2.1 Bucket Architecture

Buckets are object-oriented container constructs in which logically grouped items can be collected, stored and transported as a single unit. When one looks at the available Indian information resources, one finds the following variants and issues:

- Language (multi-lingual),
- Type of document (palm leaves to software data sets),
- Geographical constraints, being one of the largest Asian countries with many people maintaining oral knowledge as a resource without any written documents,
- One needs to take into consideration mass illiterates as users of these information objects,
- Collection, storage and retrieval is a Herculean task,
- One needs to do natural language processing for search and retrieval, etc.

In a typical library environment these things are placed in Document profile for Selective Dissemination of Information. In a similar way, these things are kept in logical containers by creating buckets, in order to disseminate the information to the users.

Similar to the information sources bucket container, there would be User's bucket container that deals with the user's requirements and interests. Typically, this would contain:

- User's interests,
- Language priorities,
- Type of documents etc.

These are logically kept in a helpful manner to match and retrieve the user's query. Here, either the metadata information will be matched for the first time query and then the protocol would allow the user to interact the respective bucket information directly through http protocol.

A single bucket can have multiple packages (different formats of information, multilingual, single language, region wise or topic). Packages can correspond to the semantics of the information (manuscript, jpg, gif, software, etc.) or can be more abstract entities such as metadata for the entire bucket, bucket terms and conditions, pointers to other buckets or packages etc. Even a single package can have multiple or several elements, which are typically different file formats of the same information. Each bucket will have a unique handle i.e. id and access methods such as terms and conditions, permission, etc.

In a similar fashion the User's bucket is also made available based on handle id. This would help connect information source bucket at regular intervals to have the latest

information. This way both buckets can be of dynamic nature and would be more active at all times.

By creating Document Bucket and User's Bucket one can establish the Bucket Matching System tool to retrieve the information on demand or on equal terms through batch processing to keep abreast of the latest information.

Thus, buckets provide a mechanism for logically grouping the various semantic data objects (manuscript, software, datasets, etc.) and various syntactic representation (formats of files).

5.2.2 STARTS architecture

For information retrieval purpose the paper addresses to use STARTS architecture. STARTS addresses three issues:

- Source metadata problem
- Query language problem
- Rank merging problem (8)

Since each Bucket is logically classified and kept, one can use STARTS protocol collection (source) metadata consisting of two files one containing the inherent metadata derived directly from the collection, including a complete word index for searching, and the other containing the contextual information providing ownership, coverage and contact information[9]. It is also to be noted that the search engine that is proposed to be used will be at source level and it should have capability to search and retrieve multilingual documents and retrieve the same according to user's requirements. There will be a well defined metadata collection schema which includes information that the search engine can query, the rules applied to the creation of the indexes to those attributes, and the types of queries that the collection can accept.

Although STARTS protocol uses Z39.50 standard for vocabulary control metadata for resource discovery, it has been found that Dublin Core will be suited for the requirement and implementation in a typical Indian environment that has been discussed above.

The proposed hybrid protocol uses typical rank merging that has been addressed in STARTS.

The drawback of this protocol is that it can be used only for textual resources available in each bucket (9). We need to take Indian language semantics while considering a search engine. Also one needs to address the non-textual resources like images etc. for resource discovery while designing the search engine.

6 Conclusion

As the WWW grows exponentially, discovery and retrieval of information sources grow more problematic, if it is not planned well. As there are data variants in collection development and retrieval aspects especially in Indian DL environment it is more so for the DL developer to meet the individual needs. The framework that has been proposed is just a beginning and it would go a long way to meet the goal and requirements of the individual user. Also, one needs to address the NLP (Natural Language Processing) issues, which we have not addressed for language processing and retrieval. Since no standards have been evolved for any of the Indian languages (fonts, phonetic solutions) for text processing like OCR and search engine, it is still more strenuous to project any concrete solution. Even when one considers universal protocol for data exchange with different DLs there is a need for standards for metdata, although Dublin Core has been proposed for its popularity and simplicity. With the growing need of such a protocol, one would anticipate a generalized protocol to emerge and facilitate the DLs design, development and implementation in an Indian environment.

Acknowledgment

Author would like to acknowledge Prof. Robert (Bob) B. Allen *rba@Glue.umd.edu*, University of Maryland, College park for his guidance and valuable suggestion throughout this project. Also sponsors USEFI, New Delhi and IIE, NY for giving opportunity to carry out this project as part of the Fulbright fellowship 1999-2000.

7 References

1. BURNETT (K), NG (KW) and PARK (S). Comparisons of the two traditions of Metada development. *JASIS*, 50(13), 1999, pp. 1209-1217.
2. CHIVERS (A) and FEATHER (J). Management of digital data: a metadata approach. *Electronic Library*, 16(6), 1998, pp. 365-371.
3. DANIEL (R), LAGOZE (C) and PAYETTE (SD). Metadata architecture for digital libraries. IN IEEE International Forum on Research and Technology Advances in Digital Libraries - ADL'98. April 22-24, Santa Barbara, California, 1998, pp. 276-288.
4. DEMPSEY (L.) and HEERY (R.). A Review of metadata: a survey of current resource description formats. 1997. *http://www.ukoln.ac.uk/metadata/DESIRE/overview/*
5. Dublin Core. 2000. *http://purl.org/DC*
6. GILL (T). Metadata and the world wide web. In Introduction to Metadata. 2000. *http://www.getty.edu/gri/standard/intrometadata*
7. GILLILAND-SWETLAND (AJ). Setting Stage. In introduction to metadata. 2000. *http://www.getty.edu/gri/standard/intrometadata*
8. GRAVONO (L) [et al.]. STARTS (Stanford Protocol Proposal for Internet Retrieval and search. 1997. *http://www-db.stanford.edu/~gravano/starts*

9. HEEKS (R). India's software industry: state policy, liberalisation and industrial development. New Delhi, Sage Publications, 1996.
10. HILL (L.L.) [et al.]. Collection metadata solutions for digital library applications. *JASIS*, 50(13), 1999, pp. 1169-1181.
11. HILLMANN (D). Using Dublin Core. 2000. *http://purl.org/DC/documents/wd/usageguide-2000716.htm*
12. MADHUSUDANA RAO (CR). Digital library for Indian heritage. Project report. 2000. *http://Raocr@tripod.com*
13. MALY (K), NELSON (ML) and ZUBAIR (M). Smart objects, dumb archives: a user-centric, layered digital library framework. *D-Lib Magazine*, March, 1999, *http://webdoc.gwdg.de/aw/d-lib/march99/maly/03maly*
14. MARCHIONINI (G). Research and development in digital libraries. *http://www.glue.umd.edu/~march/digital_library_R_and_D.html*
15. McCRAY (AT), GALLABHER (ME) and FLANNICK (MA). Extending the role of metadata in a digital library system. IN Proceedings IEEE Forum on Research and Technology Advances in Digital Library - ADL'99, Baltimore, MD, May 1999, 19-21, pp. 190-199.
16. Metadata related tools. *http://purl.org/tools/index.htm*
17. NELSON (ML). Buckets: aggregative, intelligent agents for publishing. NASA Technical Report Server. 1998. *http://techreports.larc.nasa.gov/ltrs/PDF/1998/tm/NASA-98-tm208419.pdf*
18. RAMANUJAM (P). Project proposal on Indian heritage. 1999, CDAC.
19. RUST (G). Metadata: the right approach: an integrated model for descriptive and rights metadata in E-commerce. 1998. *http://www.dlib.org/dlib/july98/rust/07rust.html*
20. SMITH (TR). Meta-information environment of digital libraries. *D-Lib magazine*. 1996. *http://www.dlib.org/dlib/july96/new/07smith.html*
21. SUTTON (SA). Conceptual design and development of a metadata framework for educational resources on the Internet. *JASIS*, 50(13), 1999, 1182-1192.
22. WEIBEL (S). Metadata: the foundations of resource description. *D-Lib Magazine*, 1995. *http://www.dlib.org/dlib/July95/07weibel.html*
23. WOODLEY (M). Crosswalks: the path to universal access? In introduction to metadata. 2000. *http://www.getty.edu/gri/standard/intrometadata*