Modern Information Retrieval

Chapter 17

Digital Libraries

Definitions
Architecture and Fundamentals
Social-Economical Issues
Software Systems and Case Studies
Research Challenges

Digital Libraries - Introduction

- Digital Libraries (DLs) are advanced and complex information (retrieval) systems, which offer many valued services besides searching and browsing, such as
  - document preservation and recommendation
  - reference services
  - selective information dissemination, among others
- All services are provided over various types of multimedia data (e.g., audio, video) in a distributed fashion

DLs vs. The Web

- Both were born in the e-Publishing era and share the spirit of open access and freedom to publish.
- But there are important differences:
  - Information in digital libraries is explicitly organized, described, and managed.
  - Input in DLs is normally more tightly controlled with a strong focus on improving quality.
  - DLs are built targeted to a particular community of users with specific information needs and tasks, and thus usually do provide more specialized, community-oriented services, but also require participation of the community for success.
  - Preservation is (or should be) a key aspect in DLs, contrasting with the low archival nature of the Web.

DLs vs. Traditional Libraries

- Main advantages:
  - DLs eliminate almost completely access and dissemination restrictions inherent to the physical world.
  - DLs stimulate profound changes in the traditional publishing methods, muddling the roles and responsibilities of authors, reviewers, editors, publishers, librarians.
  - The DL thus becomes the main channel of (direct) communication/ interaction among all players.
- Risks:
  - Absence of a clear entity responsible for preserving all the digital material being created
  - Legal issues related to intellectual property, rights management, and terms and conditions.

Defining Digital Libraries

- Different visions from different communities (e.g., researchers, practitioners) make difficult a consensual definition.
- An example of a technologically-oriented definition:
  - DLs are organized and focused collection of digital objects, including text, images, video, and audio, along with methods of access and retrieval, and for selection, creation, organization, maintenance, and sharing of the collection (Witten and Bainbridge, 1999).

Defining Digital Libraries

- An example of a more socially-oriented definition:
  - Digital libraries are organizations that provide the resources, includind the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities
- Crearly a very distinct view.

Defining Digital Libraries

- “Our” definition, inspired on the 5S (Stream, Structures, Spaces, Scenarios, and Societies) Framework:
  - DLs are complex information systems that help satisfy information needs of users (or societies of users), provide information services (that can be described through scenarios of use), and organize (through structures), present (through spaces), and communicate(through streams) information with users in usable ways.

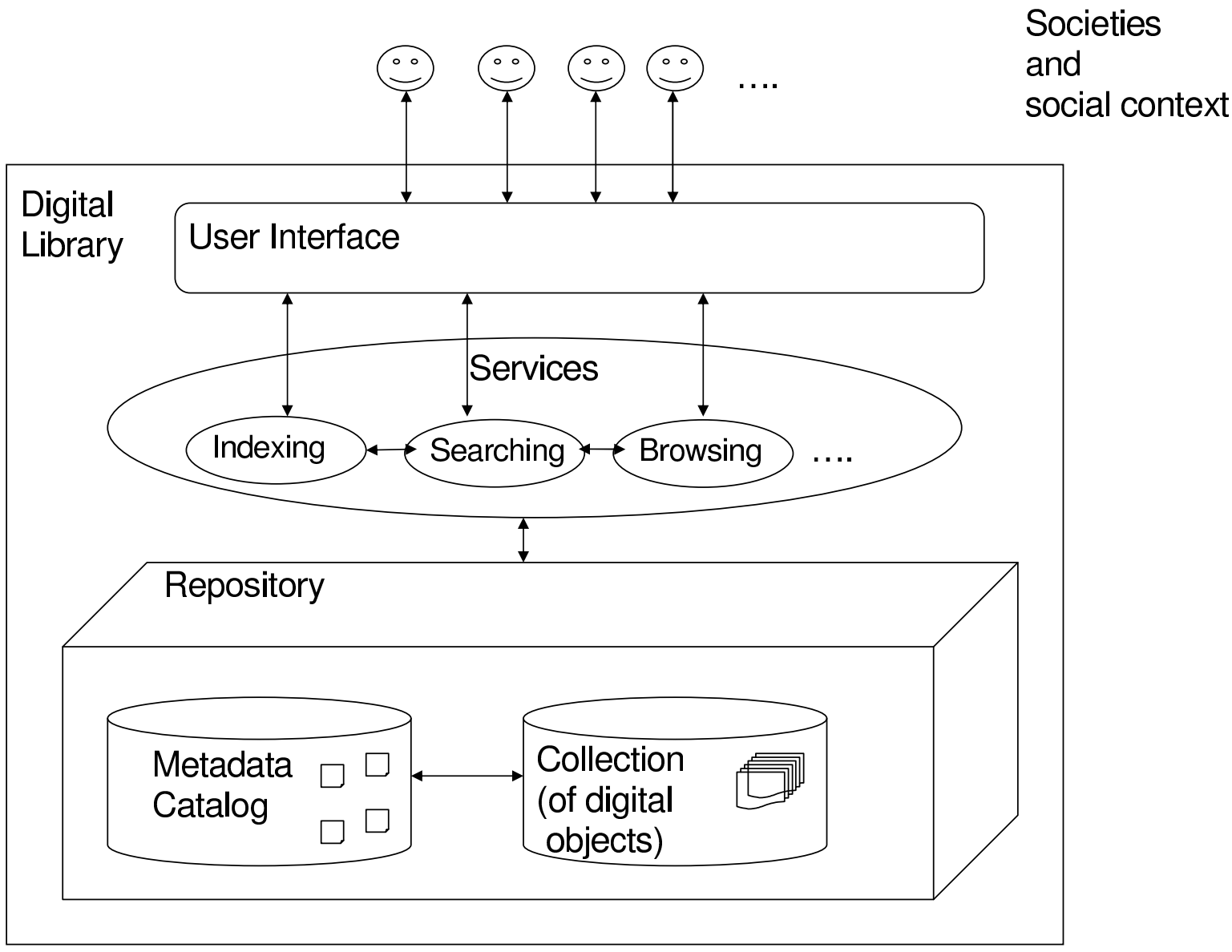
Defining Digital Libraries

- This last definition emphasizes important aspects including:
  - that multiple types of media should be supported;
  - that information is explicitly organized within a DL;
  - that rich scenarios of use, information services, and interactions may be supported;
  - that the DL has a target community (or society).

The 5S Framework for DLs

- Provides a formal foundation for DLs, by defining the basic ‘Ss’ and fundamental concepts of a ‘minimal’ DL.
  - Streams: sequences of arbitrary items (e.g., bits, characters, pixels, images) representing the content of a DL.
  - Structures: can be viewed as labeled directed graphs, which impose organization on the DL content.
  - Spaces (e.g., vector, probabilistic): used as support for services and for presentation purposes; can be seen as sets with operations that obey certain mathematical constraints.
  - Scenarios: stories that describe the behavior of services and consist of sequences of events or actions that modify states of a computation in order to accomplish a functional requirement.
  - Societies: sets of entities and activities and the relationships among them.

A General (Reference) Architecture



A General (Reference) Architecture

- A DL is comprised of a collection of digital objects (e.g., digital documents, images, etc) and a catalog of metadata records that serve either to describe, to organize, or to specify how the objects in the collection can be used and by whom.
- Ideally every object should have a corresponding metadata record in the catalog, and this record should have a specic structure defined by a schema.
- Collections and catalogs are usually stored together in a repository that provides access and management capabilities to collections and catalogs.

A General (Reference) Architecture

- *Services* to create digital objects or metadata records, to preserve content, to add value to it, and to satisfy information needs are built on top of the repositories and are used by actors in a social context.
- Services can cooperate in terms of reuse or extension of capabilities to create more advanced services from simple ones.
- Usually a digital library provides simple searching and browsing services and indexing services to support the former two as a minimal set of services.
- The user interface serves as a ”glue" to organize and display all the provided services.

Fundamentals - Digital Objects

- The main carriers of (most of) the information within DLs
- In a simple view, can be considered as composed of two itens: its content and a handle – a unique identifier
- Extensions of such simple definition may involve:
  - the inclusion of simple descriptive metadat along with the object
  - the addition of structural metadata to represent the internal organization of objects
  - the incorporation of complex behavior capable of producing different “views” or manifestations of the object (e.g., thumbnail, low and high resolution versions of a same image object or Postscript, Microsoft, Word or PDF versions of a digital document).

Fundamentals - Digital Objects

- Handles: unique digital objects' identifiers that provide a standardized way to refer to a object in order to be able to acess it from a repository, or for citation or archival purposes.
- Several syntaxes and mechanisms exist to define such identifiers:
  - Uniform Resource Identier (URI): a compact string of characters used for identifying an abstract or physical resource.
    - Uniform Resource Locator (URLs), as used in the Web, refer to the subset of URIs that identify resources via a representation of their primary access mechanism (e.g., network location).
    - Uniform Resource Names (URNs) refer to the subset of URIs that are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable

Fundamentals - Digital Objects

- Handles
  - Digital Object Identier (DOI): an open system, which conforms to URI, based on non-proprietary standards which provides a mechanism to interoperably identify and exchange intellectual property in the digital environment.
  - OpenURL: a standard syntax for transporting information (metadata and identifiers) about one or multiple resources within URLs;
  - Persistent URL (PURL): an identifier (or URL) that points to an intermediate resolution service, instead of the direct location of an Internet resource, able to redirect the identifier to an actual URL so that the client is able to complete the URL transaction in a normal fashion. While PURLs allow associating different URLs with them, the PURL itself never changes (it is *persistent*).

Fundamentals - Digital Objects

- Digitization: process of creating digital objects as surrogates of real objects (e.g., a image of an existing artefact)
- Smart (Digital) Objects:
  - Also called “buckets”, these digital objects incorporate part of the functionality of a repository such as storing and management of metadata, responsability for exporting different disseminations ot its content, and tracing of event logs.
  - Potential advantages: mobility, self-sufficiency, and repository independency.

Fundamentals - Digital Objects

- Collections:
  - In their basic incarnation can be seen as sets of digital objects.
  - May be supported by specialized supporting tools for facilitating their creation, or may be built for example by focused crawlers exploring the Web looking for specific types of documents (e.g., topic or genre-oriented).

Fundamentals - Metadata

- Metadata: data *about* digital objects; usually manisfested as “records” referring to a single object.
- It can be descriptive, structural or administrative (e.g., containing information about intellectual property rights).
- A *metadata standard* precisely defines the correct and proper use and interpretation of metadata records’ fields (also called attributes or elements) by specifying a common understanding of their meaning or semantics.
- Metadata gathering, which is usually expensive and time consuming, mainly when huge amounts of information need to be described, may be supported by automatic metadata extraction tools, a growing research area.

Fundamentals - Metadata

- Metadata Standards - Dublin Core
  - developed as a response to the complications and costs associated with complex standards such as MARC to describe Internet-based resources.
  - defines fifteen elements: seven for describing content (*title, subject, description, source, language, relation,* and *coverage*), four to deal with intellectual property issues (*creator, publisher, contributor*, and *rights*), and other four for dealing with properties of instantiations/manifestations of digital objects (*identifier, data, type*, and *format*).
  - The qualifed version of the standard defines refinements for the original 15 elements, making their meaning narrower or more restrict.

Fundamentals - Metadata

- Metadata Standards - METS (Metadata Enconding and Transmission Standard)
  - An open standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML Schema language of the World Wide Web Consortium.
  - Maintained by the Library of Congress and developed as an initiative ofthe Digital Library Federation.
  - The METS framework supports a free selection of metadata formats and avoids duplicating data that is already stored in constituent files.

Fundamentals - Metadata

- Collection-level Metadata
  - Metadata is not restricted to digital objects. Collection metadata can be used for purposes such as :
    - collection registration with the search client software that will provide access to it;
    - network discovery by providing information to network agents about what the collection contains;
    - documentation;
    - management, for instance, specifying a central point for storage of collection wide information.

Fundamentals - Metadata

- Service-level metadata
  - The W3C's Web Service Description Language (WSDL) provides mechanisms for describing Web Services in terms of the vocabulary a service understands and of the messages it is able to process as well as descriptions of specific protocol-dependent details that users must follow to access the service at concrete service end points.

Fundamentals - Repositories

- Repositories are responsible for storing collections of digital objects and providing basic methods for depositing and retrieving specific objects based on their handles.
- Additional features such as security, and most importantly, a repository access protocol for remote and distributed access may also be provided.
- *Institutional repositories* are responsible for storing and archiving the complete intellectual production of a particular institution or consortium of institutions, in most cases educational ones (e.g., universities), for long-term preservation, access, and distribution.
- Two other important issues related to repositories are: *interoperability* and *preservation*.

Fundamentals - Interoperability

- Refers to the capability of distributed DL systems to work together to achieve common goals such as distributed searching or browsing.
- Solutions differ in the in the amount of standardization or effort required from each DL component.
- The three main approaches are *federated services*,*harvesting*, and *gathering*.

Fundamentals - Interoperability

- Federated Services
  - Here, a group of organizations decide that their services will be built according to a number of agreed upon specifications, normally selected from formal standards. The work of forming a federation is the effort required by each organization to implement and keep current with all the agreements.
  - Examples: Federations based on Z39.50 (a comprehensive protocol to allow client/server intercommunication between retrieval systems) or SRU/SRW (based on Z39.50 but less complex, is built using Web standards (SOAP, HTTP, XML)).
  - A federated service is as vulnerable as its weakest component, in terms of performance and reliability.
  - Creating large federations implies in costs to implement the standards and agreements and keep current with any changes.

Fundamentals - Interoperability

- Harvesting
  - Based on the idea of creating looser groupings of digital libraries when compared to federations.
  - The participants should make some small efforts to enable some basic shared services, without specifying a complete set of agreements.
  - Example – Open Archives Initiative (OAI): promotes the use of Dublin Core as a standard metadata format and defines a simple standard metadata harvesting protocol. Metadata from DLs implementing the protocol can be harvested to central repositories upon which DL services can be built.
  - Small local repositories may still lack staff resources to install and maintain a OAI server. Others will not take any active steps to open their contents at all.

Fundamentals - Interoperability

- Gathering
  - Only solution when there is no formal cooperation.
  - Best exemplified by Web Crawling.
  - Requires minimum effort (e.g., set up a web server) but implies in poorer quality of services when compared to previous approaches.

The Open Archives Initiative (OAI)

- Considered by many one of the most important developments in the DL field.
- Based on a clear separation between Data Providers and Service Providers
  - Data providers: use the OAI technical framework as a means of exposing metadata about their content.
  - Service providers: harvest metadata from data providers using the OAI protocol and use the metadata as the basis for building value-added services.
- All OAI data providers supply metadata in a common format: the Unqualified Dublin Core Metadata Set.

The Open Archives Initiative (OAI)

- Community-specific descriptions, or metadata specificity, is addressed in the technical framework by supporting for parallel metadata sets.
- Two types of selective harvesting are allowed:
  - By date: requests may contain a date range for harvesting, that may be total (between two dates) or partial (either only a lower bound or an upper bound) based on the record's datestamp.
  - By sets: an optional, possibly hierarchically organized groups of items. The actual meaning of the sets is not specified by the OAI infrastructure.

The Open Archives Initiative (OAI)

- The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) contains six verbs:
  - GetRecord – used to retrieve an individual (metadata) record from an item in a repository. Required arguments specify the identifier (or handle) of the requested record and the format of the metadata that should be included in the record.
  - Identify – used to retrieve information about a repository including: 1) a human readable name for the repository; 2) the base URL of the repository; 3) the version of the OAI protocol supported by the repository; 3) the email address of the administrator of the repository.

The Open Archives Initiative – Verbs

- OAI-PMH contains six verbs:
  - ListIdentifers – used to retrieve the identifiers of records that can be harvested from a repository. Optional arguments permit selectivity of the identifers - based on their membership in a specific set in the repository or based on their modification, creation, or deletion within a specifc date range.
  - ListMetadataFormats – used to retrieve the metadata formats available from a repository. An optional argument restricts the request to the formats available for a specific record.
  - ListRecords – used to harvest records from a repository. The same selectivity criteria of the ListIdentifers verb can be used.
  - ListSets – used to retrieve a respository set structure.

Preservation and Archiving

- Software and hardware obsolescence and computer media degradation are factors that put all types of digital content at risk of not being available for interested users in the not so distant future.
- As more and more digital information is created, attention must be paid to what information should be preserved and how it can be preserved in a manner that is economical and effective.
- It is pervasive view in the community that repositories and archives are responsible for preserving the content they hold.
  - These archives should guarantee that the information is reliably authentic and intelligible.
  - They should transparently expose their procedures and practices in order to increase the public perception that an archive has correctly executed sound preservation practices.

Preservation and Archiving

- Five main approaches for digital preservation:
  - Migration – Transforming from one digital format to another format, normally a successive subsequent one (e.g., from JPEG to JPEG 2000).
  - Emulation – Re-creating the original operating environment by saving the original programs and or creating new programs that can emulate the old environment.
  - Wrapping – Packaging the object to be preserved with enough human readable metadata to allow it to be decoded in the future. This idea is explored in a series of works by Gladney who coined the term *trustworthy digital objects*: a digital object encapsulated with metadata describing its origins, cryptographic sealing, webs of trust for public keys rooted in a set of respected institutions, and specific ways of managing information identifiers.

Preservation and Archiving

- Five main approaches for digital preservation:
  - Refreshing – Copying the stream of bits from one location to another, whether the physical medium is the same or not.
  - Replication – Making enough copies of the data. This is the major approach advocated by the LOCKSS project. The LOCKSS project has developed and deployed in a world-wide test a peer-to-peer system for preserving access to digital material. It consists of a large number of independent, low-cost, persistent Web caches that cooperate to detect and repair damage to their content by using sophisticated voting schemes.
- The ideal solution should be a combination of two or more of these techniques, considering the particularities (e.g., cost, operational, and technical settings) of each case.

Fundamentals – Services

- The main channel of access to DLs.
- Through services information is created, discovered, enriched, accessed and ultimately used in digital libraries.
- DL services can be characterized and understood by analyzing inputs (what they consume) from either users (patrons, actors) or from other services, and their outputs (what they produce). This is important, for example, when services are accessible only in a black box mode without knowledge of their internal procedures or components.
- Such an analysis helps to understand how a service is used, the requirements it attends, and its boundaries.

DL Services Taxonomy

Infrastructure Services			*Information Satisfaction Services*
Repository-Building		*Add-Value*	
Creational	Preservational		
Acquiring	Conserving	Annotating	Binding
Authoring	Converting	Classifying	Browsing
Cataloging	Copying/Replicating	Clustering	Customizing
Crawling (focused)	Migrating	Evaluating	Disseminating
Describing	Translating (format)	Extracting	Expanding (query)
Digitizing		Indexing	Filtering
Harvesting		Linking	Recommending
Submitting		Logging	Requesting
		Measuring	Searching
		Rating	
		Reviewing (peer)	
		Surveying	
		Training (classifier)	
		Translating (language/format)	
		Visualizing	

Fundamentals - Services

- Other important issues regarding services include:
  - Usability: the ease with which actors can and use DL services, mainly through their user interfaces.
  - Accessibility: also a social issue, refers to the degree to which a service is available to as many people as possible, mainly considering people with disabilities.
  - Log Analysis: can be a primary source of knowledge about how digital library patrons actually use DL systems and services and how these behave while trying to support user information seeking activities.
    - They allow evaluation assessment, and open opportunities to improvements and enhanced new services.
    - Proposals for XML-based log standards for digital libraries, which carry more information than traditional web server logs have been suggested.

Social-Economical Issues

- Some of the most difficult challenges in the area of Digital Libraries are not technical, but involve social and economical issues and practices.
- Social Issues
  - The most prominent ones include:
    - aspects of cooperation and collaboration (including sharing of information);
    - social networks;
    - cultural heritage and preservation;
    - digital divide and internationalization.

Social-Economical Issues

- Cooperation and Collaboration
  - Collaborative production of content (in some cases competitive) as best exemplied by the Wikipedia project, is key to guarantee the sustainability of long-term projects by exploring voluntary and community-regulated production of an intellectual work.
  - A community of volunteers contribute project components, in a coordinated way, but without a traditional hierarchical organization or financial compensation, and there exists some process to combine them to produce a unified work.
  - This form of collaborative production of content has arisen because the Internet has lowered certain communication and collaboration barriers, allowing the creation of large and complex projects.

Social-Economical Issues

- Social Networks
  - Informal [communities] of collaborators, colleagues, or friends with shared interests, these networks connect people together through implicit or explicit interests and communications or interactions.
  - Best exemplified by relationship sites such as Facebook, Orkut, Friendster, and MySpace.
  - Have been explored in a few DL projects in applications that vary from the aforementioned collaborative efforts, to the disambiguation of authorship of scientific articles by exploring the co-authors relationships, the provision of recommendation services, and the integration of DLs into specific community networks.

Social-Economical Issues

- Preservation
  - The most difficult social issues involved with this topic involve establishing a public awareness of the enormous dangerous of not considering preservation issues in the information life cycle with the consequences of loosing most of the (digital) knowledge being produced nowadays as well as the cultural heritage that is being currently digitized.
  - Moving from this consciousness to a preservation culture in which preservation aspects are embedded within the chain of production and consumption of knowledge is an enormous challenge that involves:
    - defining responsibilities;
    - changing established practices and methods;
    - finding resources to deal with additional costs;
    - assuring authenticity of the preserved information, among many others.

Social-Economical Issues

- Economical Issues – involve aspects such as:
  - security: authorization, authentication, watermarks;
  - legal aspects: terms and conditions,patents, trademarks, copyright, intellectual property rights, digital rights management;
  - publishing: self-archiving, cataloguing costs, open collections);
  - sustainability.

Social-Economical Issues

- Economical Issues
  - Sustainability: most effective approaches exploit collaboration and envolviment of the target community as previously discussed.
  - Rights management: include legal and technical systems that protect the right of individuals to make exclusive use of the expressed form of the products of their ideas or intellect.
    - One example is copyright: the right to make copies of an original creation, usually for a limited period of time.
    - Traditional rights management benefited from the material physicality. The change to the digital arena and the easiness with which digital material can be copied and transmitted has brought many challenges regarding the protection of these rights.
    - Digital Rights management is a term that encompasses not only issues of security and encryption but also the description, identification, trading, protection, monitoring and tracking of the all forms of rights held over digital material.

Social-Economical Issues

- Economical Issues
  - Publishing issues – Open Acess Movement: advocates free access to research publications trying to undercome problems with the current process of scientific communication including:
    - the imbalance between publishers prices for journal licenses, which grow much faster than inflation, and the low budget of research and university libraries;
    - the impossibility of authors to promote and share their own work with peers and therefore obtain the necessary scientific acknowledgment, due to copyright transfers to the publishers;
    - the enormous delays between submission and the actual publication of the work in the literature.

Social-Economical Issues

- Economical Issues
  - Publishing issues – Two concepts are important in this context:
    - Pre-prints: pre-publication versions of scientific articles, that facilitate the distribution of such items, thus allowing more immediate feedback to the authors.
    - Self-Archiving: allow researchers to easily archive their own work and share results and publications with their peers, thus saving costs by involving the interested community.

Digital Library Software Systems

- Software systems that give support to the creation of DLs.
- Most of them can be freely downloaded and installed locally, providing the support to the construction of many DLs by the target communities.
- Many of the current and most used DL software systems were developed as research prototypes or in a partnership with some enterprises.

Digital Library Software Systems

- Greenstone (University of Waikato): provide tools for for facilitating the process of assembling and enriching digital library collections, building several browsing indexes from metadata, and prooessing different types of digital content.
- Eprints (University of Southampton): designed to promote self-archiving of scientific research materials, it is a leading software in the implementation of Institutional Repositories with a large and growing installation base around the world.

Digital Library Software Systems

- Dspace (MIT Libraries & Hewlett-Packard): focused on developing and deploying solutions for long term preservation of digital research and educational material, this system includes a workflow module that supports the submission process whose policies can be adapted for different communities and collections.
- Fedora (Cornell University & University of Virginia Library): developed for storing, managing and disseminating complex digital objects with several interconnected types of content and metadata, along with methods for exporting several (dynamic) disseminations, as well as relationships among these objects.

Digital Library Software Systems

- ODL (Virginia Tech): advocates a componentized approach for the development of digital libraries. The functionality of the digital library is implemented in several software components, which allows for reusability and modularity. Each component is implement as an Open Archive and inter-component communication is implemented using extensions of OAI-PMH.
- 5S Suite (Virginia Tech): more a proof-of-concept than a real working system, the 5S suite of tools takes a model-driven approach for building and generating DL applications. Organized around 5S, includes among others, a domain specific modeling language (5SL), a graphical modeling tool (5SGraph), and a code generator which also exploits the ODL components (5SGen).

DL Case Studies

- The Networked Digital Library of Theses and Dissertations (NDLTD)
  - A global network of DLs focused on electronic theses and dissertations, illustrates a very successful digital library case that had initially no central governing body and almost no budget.
  - The idea to promote sustainability was to involve as many institutions as possible worldwide, by showing them that all involved in the process would benefit somehow from participating in the initiative.
  - University libraries would see an enormous economical benefit from not having to archive paper copies of the defended theses and dissertations.

DL Case Studies

- The Networked Digital Library of Theses and Dissertations (NDLTD)
  - The Graduate School could streamline and lower the costs of the submission and handling of ETDs by having an electronic workflow supporting these activities.
  - Professors and students would have more visibility to their work. Students would also learn more about the production of electronic documents.
  - In those institutions which require students to catalog their own works, they would also gain knowledge that would benefit their future careers as scholars.

DL Case Studies

- The Networked Digital Library of Theses and Dissertations (NDLTD)
  - Institutions would open the doors for the knowledge they produce, which is almost unknown by the general public.
  - NDLTD has hundreds of institutions as members spread in the five continents, including consortia such as the British Library and Unesco. Unesco has also sponsored the creation of the ETD Guide to help institutions to produce their own ETD programs.
  - The NDLTD Union Catalog contains almost 1,9 million entries (as of January, 2011).
  - Several searching, browsing, and clustering services are built on top of their union catalog (harvest via OAI-PMH).
  - The initiative has also promoted its own standards for cataloguing and sharing ETD metadata.

DL Case Studies

- The ETANA-DL Archaelogical Digital Library
  - A domain-specific archaeological digital library aimed at collecting, recording, integrating and preserving archaeological data gathered during archaeological surveys and excavations.
  - The current version of ETANA-DL aggregates data collected at several archaeological sites from the Middle East and beyond. These data include, for example, figurine pho- tos, information about excavated seeds and bones, drawings and stratographic maps of excavated units.
  - ETANA-DL has services for archaeologists (e.g., content-based image retrieval) and the general public.
  - The experience used in ETANA-DL can serve as a model for the creation of other domain specific digital libraries (e.g., ecological/biodiversity, educational).

Research Challenges

- Many research challenges were already discussed, including:
  - information (mainly metadata) extraction;
  - new models for (complex) digital objects;
  - repository management and scalability;
  - interoperability;
  - digital preservation;
  - social-economical issues;
  - DL architectures.
- A few other ones not previously discussed are covered next.

Research Challenges

- Evaluation
  - Evaluation of digital libraries is difficult due to the many aspects involved and to the lack of consensus about how to take all these aspects into consideration conjointly.
  - A few evaluation models have been proposed, each one considering different criteria.
  - All after all, the only consensus is that agreement on common evaluation practices and methods will take considerable time as DL research is still at an early stage and wide-scale acceptance and employment of DLs is only in the beginning.
  - Evaluation though is crucial for the evolution and acceptance of this type of system.

Research Challenges

- Integration: means hiding distribution and heterogeneity, while at the same time enabling and making visible component autonomy (at least to some degree).
  - The inability to seamlessly and transparently access knowledge across DLs is a major impediment to knowledge sharing.
  - The goal of DL integration then is to utilize various autonomous DLs in concert to provide knowledge from such island-DLs.
  - The needs for DL integration are well known, and better known than the solutions.

Research Challenges

- Creation of a Reference Model for DLs –the lack of a consensual reference model for digital libraries, which lays the foundations for the field as a whole, has slowed down progress by making efforts very hard to compare and integrate and results to be shared and reused. The 5S framework and the DELOS reference model are steps towards this goal.
- Citation management in DLs – involve aspects such as:
  - information extraction methods to parse the correct components of the citations in any given format;
  - data cleaning to correct mistakes, such as assignment of improper authorship or splitting of a researcher's production due to the use of multiple names in publications, a problem known as name disambiguation; and
  - removal of duplicates, mainly after data integration or data input tasks, a problem known as record duplicate detection or deduplication.

Research Challenges

- Personalization: adaptation of content and services to an individual or person preferences and idiosyncrasies.
  - Novel techniques to enable, for example, capturing of the attention that a user spends on specific resources in a specific context and identifying and mining patterns of user behavior need to be developed.