1-a (10-c): Digital Library Curriculum Development
Module: Conceptual Frameworks, Models, Theories, and Definitions
Date: May, 11 2011

1. **Module name:** Conceptual frameworks, models, theories, and definitions

2. **Scope:** Introduction to several conceptual models characterizing the DL domain (Digital Libraries Reference Model-DLRM, 5S, DELOS Classification and Evaluation Scheme, CIDOC Conceptual Reference Model, DOLCE-based Ontologies for Large Software Systems)

3. **Learning objectives:** Students will be able to:
   - Students will be provided with a high level yet comprehensive knowledge of several conceptual frameworks and models.
   - Students will be provided with a unifying and extended terminology.
   - Students will be provided with an overall scheme helping to classify further readings.

4. **5S characteristics of the module:**
   - Spaces: Reference is provided in an information space, physical or online.
   - Scenarios: Reference is an info seeking technique employed by users in specific situations / contexts / anomalous states of knowledge.
   - Societies: Reference is provided by a community of answerers (usually librarians and/or subject experts) and used by communities of users.

5. **Level of effort required:**
   - Prior to class: 4 hours for readings
   - In class: 2.5 hours
   - Exercises: 3 hours

6. **Relationships with other modules:**
   - Introductory, can be read independently from other modules. If, however, read at the beginning, it can offer a better understanding of other modules.

7. **Prerequisite knowledge required:**
   N/A

8. **Introductory remedial instruction:**
   N/A

9. **Body of knowledge:**
The Digital Libraries Reference Model

What is a Reference Model?

A reference model is an abstract framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment.

A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist.

A reference model is not directly tied to any standards, technologies or other concrete implementation details, but it does seek to provide a common semantics that can be used unambiguously across and between different implementations.

Objective and structure of the Digital Libraries Reference Model

Sets the foundations and identifies the cornerstone concepts within the universe of Digital Libraries, facilitating the integration of research and proposing better ways of developing appropriate systems. Reflecting the structure of the DL universe the DLRM is segregated into six (6) sub-domains comprising interrelated concepts and terms i.e. the Content, User, Functionality, Quality, Policy and Architecture domains. Concepts from each of these sub-domains are materialized in nearly every existing DL. These concepts are further analyzed next.

In terms of document structure the DLRM consists of 3 parts:
- Digital Library Manifesto
- Digital Library Reference Model in a Nutshell
- Digital Library Reference Model Concepts & Relations.

The Reference Model fundamentals

A Three-tier Framework

a. Digital Library (DL)

An organization, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content, of measurable quality and according to codified policies.

b. Digital Library System (DLS)

A software system that is based on a defined (possibly distributed) architecture and provides all functionality required by a particular Digital Library. Users interact with a Digital Library through the corresponding Digital Library System.
c. Digital Library Management System (DLMS)

A generic software system that provides the appropriate software infrastructure both (i) to produce and administer a Digital Library System incorporating the suite of functionality considered fundamental for Digital Libraries and (ii) to integrate additional software offering more refined, specialized or advanced functionality.

The Constituent Domains

The *Digital Library Domain*, which comprises all the elements needed to represent the three systems of the DL universe, is divided into two main classes: *DL Resource Domain* and *Complementary Domain*. The DL Resource Domain contains elements identified as ‘first class citizens’ in modeling the Digital Library universe. The classification is as follows:

**Content**

The *Content* concept encompasses the data and information that the Digital Library handles and makes available to its users. It is composed of a set of information objects organized in collections. Content is an umbrella concept used to aggregate all forms of information objects that a Digital Library collects, manages and delivers. It encompasses the diverse range of information objects, including such resources as objects, annotations and metadata. For example, metadata have a central role in the handling and use of information objects, as they provide information critical to its syntactical, semantic and contextual interpretation.

**User**

The *User* concept covers the various actors (whether human or machine) entitled to interact with Digital Libraries. Digital Libraries connect actors with information and support them in their ability to consume and make creative use of it to generate new information. User is an umbrella concept including all notions related to the representation and management of actor entities within a Digital Library. It encompasses such elements as the rights that actors have within the system and the profiles of the actors with characteristics that personalize the system’s behavior or represent these actors in collaborations.

**Functionality**

The *Functionality* concept encapsulates the services that a Digital Library offers to its different users, whether classes of users or individual users. While the general expectation is that DLs will be rich in capabilities and services, the bare minimum of functions would include such aspects as new information object registration, search and browse. Beyond that, the system seeks to manage the functions of the Digital Library to ensure that the functions reflect the particular needs of the Digital Library’s community of users and/or the specific requirements relating to the Content it contains.

**Quality**
The *Quality* concept represents the parameters that can be used to characterize and evaluate the content and behavior of a Digital Library. Quality can be associated not only with each class of content or functionality but also with specific information objects or services. Some of these parameters are objective in nature and can be measured automatically, whereas others are subjective in nature and can only be measured through user evaluations (e.g. focus groups).

**Policy**
The *Policy* concept represents the set or sets of conditions, rules, terms and regulations governing interaction between the Digital Library and users, whether virtual or real. Examples of policies include acceptable user behavior, digital rights management, privacy and confidentiality, charges to users, and collection delivery. Policies belong to different classes; for instance, not all policies are defined within the DL or the organization managing it. The policy supports the distinction between extrinsic and intrinsic policies. The definition of new policies and re-definition of older policies will be a feature of digital libraries.

**Architecture**
The *Architecture* concept refers to the Digital Library System entity and represents a mapping of the functionality and content offered by a Digital Library on to hardware and software components.

The *Complementary Domain* contains all the other domains, which, although they do not constitute the focus of the digital libraries and can be inherited from existing models, are nevertheless needed to represent the systems. This concept serves as a placeholder for domains different from those identified as ‘first class citizens’ and as a hook for future extensions of the model. It includes concepts such as:

- **Time Domain** (i.e. concepts and relations needed to capture aspects of the time sphere such as time periods and intervals);
- **Space Domain** (i.e. concepts and relations needed to capture aspects of the physical sphere such as regions and locations);
- **Language Domain** (i.e. concepts and relations needed to capture aspects of the method of communication, either spoken or written, consisting of the use of words in a structured and conventional way).

**Players acting in the DL universe**

- *The DL End-Users are the ultimate clients the Digital Library is going to serve.* They exploit the DL functionality for providing, consuming, and managing the DL Content as well as some of its other constituents. They perceive the DL as a state full entity that serves their functional needs. DL end-users may be further partitioned into • Content Creator • Content Consumer • Librarian
- *The DL Designers are the organizers and orchestrators of the Digital Library from the application point of view.* They exploit their knowledge of the application semantic domain to define, customize, and maintain the Digital Library so that it is aligned with the information and functional needs of its end-
users. To perform this task, they interact with the DLMS providing functional and content configuration parameters.

- **The DL System Administrators are the organizers and orchestrators from the physical point of view.** They select the software components necessary to create the Digital Library System needed to serve the required DL and decide where and how to deploy them. They interact with the DLMS by providing architectural configuration parameters, such as the selected software components, the hosting nodes, and the components allocation.

- **The DL Application Developers are the implementers of the software parts needed to realize the Digital Library.** They develop the software components of DLMSs and DLSs, realizing the necessary functionality.

### The 5S Framework: Streams, Structures, Spaces, Scenarios and Societies

The 5S Framework is the result of an activity aimed at defining digital libraries in a rigorous manner. It is based on five fundamental abstractions, namely Streams, Structures, Spaces, Scenarios and Societies. These five concepts are informally defined as follows:

- **Streams** are sequences of elements of an arbitrary type (e.g. bits, characters, images) and thus they can model both static and dynamic content. Static streams correspond to information content represented as basic elements, e.g. a simple text is a sequence of characters, while a complex object like a book may be a stream of simple text and images. Dynamic streams are used to model any information flow and thus are important for representing any communication that takes place in the digital library. Finally, streams are typed and the type is used to define their semantics and application area.

- **Structures** are the way through which parts of a whole are organized. In particular, they can be used to represent hypertexts and structured information objects, taxonomies, system connections and user relationships.

- **Spaces** are sets of objects together with operations on those objects conforming to certain constraints. This type of construct is powerful and, as suggested by the conceivers, when a part of a DL cannot be well described using another of the 5S concepts, space may well be applicable. Document spaces are the key concepts in digital libraries. However, spaces are used in various contexts – e.g. indexing and visualizing – and different types of spaces are proposed, e.g. measurable spaces, measure spaces, probability spaces, vector spaces and topological spaces.

- **Scenarios** are sequences of events that may have parameters, and events represent state transitions. The state is determined by the content in a specific location but the value and the location are not investigated further because these aspects are system dependent. Thus a scenario tells what happens to the streams in spaces and through the structures. When considered together, the scenarios describe the services, the activities and the tasks representing digital library functions. DL workflows and dataflows are examples of scenarios.

- **Societies** are sets of entities and relationships. The entities may be humans or software and hardware components, which either use or support digital library
services. Thus, society represents the highest-level concept of a Digital Library, which exists to serve the information needs of its societies and to describe the context of its use.

We can relate the 5S to some of the aims of a Digital Library:

- Societies define how a Digital Library helps in satisfying the information needs of its users.
- Scenarios provide support for the definition and design of different kinds of services.
- Structures support the organization of the information in usable and meaningful ways.
- Spaces deal with the presentation and access to information in usable and effective ways.
- Streams concern the communication and consumption of information by users.

These concepts are of general purpose and represent low-level constructors. Using these concepts, Gonçalves et al. introduced a DL ontology. In this ontology, the different Ss are defined starting from basic mathematical concepts, such as graph or function, and are then combined and used to introduce the specific concepts that characterize the Digital Library universe. For example, the concept of digital object is defined in terms of the streams and structures that constitute it and, in turn, is used for introducing the concept of collection. In accordance with this framework, Gonçalves et al. define a minimal Digital Library as a quadruple (R,Cat,Serv,Soc) where:

- R is a repository, a service encapsulating a family of collections and specific services (get, store and del) to manipulate the collections;
- Cat is a set of metadata catalogues for all collections in the repository;
- Serv is a set of services containing at least services for indexing, searching and browsing; and
- Soc is a society.

On top of this, a framework aimed at arranging the concepts and identifying the relationships between them has been proposed.

**Comparison: Reference Model and 5S Framework**

There is a correspondence between the area covered by the 5S framework and the Reference Model: 5S basically covers what in the Reference Model have been called Content, Functionality and User main concepts; the Quality main concept has been addressed separately in the 5S Quality model, while the Policy main concept has scarcely been dealt with in the 5S framework. Moreover, the degree of detail in the different areas can vary, since in some areas the 5S framework introduces very fine-grained concepts while in other areas it adopts a more high-level approach; similar considerations also hold for the Reference Model.

Besides the above difference, it is also important to note the similarity arising around the notion of Information Object, termed digital object in the 5S framework. This probably indicates that the information object concept has been investigated more and is probably better understood than other elements constituting the Digital Library universe.
The DELOS Classification and Evaluation Scheme

The DELOS Working Group dealing with the evaluation of digital libraries problem proposed a model that is broader in scope than the one usually adopted in the evaluation context. The aim is to be able to satisfy the needs of all DL researchers, either from the research community or from the library community. This group started from a general-purpose definition of Digital Library and identified three non-orthogonal components within this digital library domain: the users, the data/collection and the chosen system/technology. These entities are related and constrained by means of a series of relationships, namely:

1. The definition of the set of users predefines the range and content of the collection relevant and appropriate for them.
2. The nature of the collection predefines the range of technologies that can be used.
3. The attractiveness of the collection content with respect to the user needs and the ease of use of the technologies by these users determine the extent of usage of the DL.

By relying on these core concepts and relationships, it is possible to move outwards to the DL Researcher domain and create a set of researcher requirements for a DL test-bed. Recently, this model has been enriched by focusing on the inter-relationships between the basic concepts, i.e. the User–Content relationship is related to the usefulness aspects, the Content–System relationship is related to the performance attributes, while the User–System is related to usability aspects. For each of these three aspects, techniques and principles for producing quantitative data and implementing their evaluation have been introduced. The Reference Model addresses similar issues through the Quality domain (please check section 9.1). While the evaluation framework takes care of identifying the characteristics of the DL systems to be measured and evaluated, the Digital Library Reference Model introduces this notion at the general level of Resource, i.e. each Resource is potentially subject to various judgment processes capturing different perspectives.

The CIDOC Conceptual Reference Model

The CIDOC Conceptual Reference Model (CRM) is an initiative whose goal is to provide a model, i.e. a formal ontology, for describing implicit and explicit concepts and relationships needed to describe cultural heritage documentation. This activity started in 1996 under the auspices of the ICOM-CIDOC Documentation Standard Working Group and since December 2006 it has been an official ISO standard (ISO 21127:2006).

It consists of 81 classes, i.e. categories of items sharing one or more common traits, and 132 unique properties, i.e. relationships of a specific kind linking two classes. Moreover, classes as well as properties are organized in a hierarchy through the ‘is a’ relationship.

The CIDOC reference model classifies the rest as the CRM Entity, i.e. the class comprising all things in the CIDOC universe and the Primitive Value class, i.e. the class representing values used as documentation elements (Number, String and Time Primitive). This second class is not elaborated further. The entities of the CIDOC
universe are further classified in *Temporal Entity*, i.e. phenomena and cultural manifestations bounded in time and space; Persistent Item, i.e. items having a persistent identity; *Time-Span*, i.e. abstract temporal extents having a beginning, an end and a duration; *Place*, i.e. extents in space in the pure sense of physics; and Dimension, i.e. quantifiable properties that can be approximated by numerical values.

The *Persistent Item* class can be compared to our notion of Resource as univocal identified entity (Resource Identifier). It is further specialised to form a hierarchy. Thing is the direct subclass and represents usable discrete, identifiable instances of persistent items documented as single units. At this point a complex hierarchy of things classes is introduced. In this hierarchy three classes need to be further explained, namely *Conceptual Object, Information Object* and *Collection*. A *Conceptual Object* is defined as ‘non-material product of our minds, in order to allow for reasoning about their identity, circumstances of creation and historical implications’. It shares many commonalities with the IFLA-FRBR concept of Work, while its counterpart in the Digital Library Reference Model is the *Information Object*.

The CIDOC-CMR *Information Objects* are defined as ‘identifiable immaterial items, such as poems, jokes, data sets, images, texts, multimedia objects, procedural prescriptions, computer program code, algorithm or mathematical formula, that have an objectively recognisable structure and are documented as single units’. The CIDOC *Information Object* concept falls within the concept of *Information Object* of the Digital Library Reference Model. The CIDOC model takes care of complex *Information Objects* through the ‘is composed of’ property as well as of rights ownership through the linking between Legal Object and Right. Collection is defined as ‘aggregation of physical items that are assembled and maintained by one or more instances of Actor over time for a specific purpose and audience, and accounting to a particular collection development plan’. Thus, differing from the Digital Library Reference Model, the CIDOC-CRM only refers collections to physical instantiation of such *aggregative mechanism*.

*Actor*, i.e. people who individually or as a group have the potential to perform actions of which they can be deemed responsible, is introduced as a specialisation of the Persistent Item class. This concept presents many commonalities with the one introduced in the Digital Library Reference Model and presented in Section 9.1. Another specialisation of the *Persistent Item* class is *Appellation*, i.e. any sort of identifier that can be used to identify specific instances of all the classes. The two models dedicate a different effort to modelling this aspect. While the Digital Library Reference Model introduces the concept of Resource Identifier without specialising it, the CIDOC-CRM introduces many specialisations ranging from *Object Identifier* to *Address, Title and Date*.

Finally, the CIDOC-CRM captures also aspects related to the notion of Functionality. In fact, even if its goal is to provide an ontology for modelling cultural heritage information, some of its classes aim at capturing the history and evolution of such information and thus can be considered as a sort of Function to which objects/information have been subjected. In particular, the role of the *Activity* class is to comprise ‘actions intentionally
carried out by instances of Actor that result in changes of state in the cultural, social, or physical systems documented’.

DOLCE-based Ontologies for Large Software Systems

DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) is a foundational ontology developed to capture the ontological categories underlying natural language and human common sense. By relying on the basic constructs it identifies, a framework of a set of ontologies for modelling modularisation and communication in Large Software Systems has been developed. This framework consists of three ontologies:

- the Core Software Ontology (CSO);
- the Core Ontology of Software Components (COSC); and
- the Core Ontology of Web Services (COWS).

The first of these provides foundations for describing software in general. In particular, it introduces the notions of ‘Software’ and ‘ComputationalObject’, which represent respectively the encoding of an algorithm and the realisation of a code in a concrete hardware. These notions are similar to the Software Component and Running Component notions envisaged by the Reference Model. In addition, the CSO ontology introduces concepts borrowed from the object-oriented paradigm such as ‘Class’, ‘Method’ and ‘Exception’, which from the Reference Model point of view are considered fine-grained and relegated to Concrete Architecture models. This ontology contains also the concepts for dealing with access rights and policies. In particular, by relying on the ‘Descriptions & Situations’ constructs of the DOLCE ontology, the concepts of ‘PolicySubjects’ (which can be ‘Users’ or ‘UserGroups’), ‘PolicyObjects’ (which can be ‘Data’) and ‘TaskCollections’ (set of ‘ComputationalTasks’) are introduced. The former two aspects are captured in a general manner by the Reference Model through the relationship between the Resource and the Policy concepts, i.e. <regulatedBy>, and through the concept of Role (and Resource Set) with respect to the intuition behind ‘TaskCollections’.

The Core Ontology of Software Components provides concepts needed to capture software components related aspects like libraries and licenses, component profiles and component taxonomies. The notion of ‘SoftwareComponent’ (having a ‘Profile’ aggregating knowledge about it) is the main entity in this ontology and it is formalised as a ‘Class’ that conforms to a ‘FrameworkSpecification’ (a set of ‘Interfaces’). Moreover, the notion of ‘SoftwareLibrary’ and ‘License’ completes the scenario by introducing notions for supporting the automatic check of conflicting libraries and incompatible licenses. The similarities with the set of concepts captured by the Reference Model Architecture Domain are evident. However, it is important to notice that the way the dependencies between the various components are captured by the Reference Model enables it to be more flexible with respect to this point.

The Core Ontology of Web Services reuses all the other ones to establish a well-founded ontology for Web Services. This is a very specific ontology that captures the component-
oriented approach in terms of standards for protocols (SOAP) and descriptions (WSDL). The other interesting feature is the explicit introduction of the ‘QualityOfService’ parameters, which in the case of the Reference Model are captured through the general relationship, i.e. <hasQuality>, between a Resource and its Quality Parameters.

10. Resources


11. Concept map
The Reference Model describes the Digital Library Universe by Concept Maps i.e. graphical tools for organizing and representing knowledge in terms of Concepts and Relationships. Concept maps are graphical tools for organizing and representing knowledge in terms of concepts (entities) and relationships between concepts to form propositions. Concepts are used to represent regularity in events or objects, or records of events or objects. Propositions are statements about some objects or events in the universe, either naturally occurring or constructed. Propositions contain two or more concepts connected using linking words or phrases to form a meaningful statement. In the graphical representation, concepts are inscribed in circles or boxes, while propositions (proposition connectors) are represented as (directed) lines connecting concepts, labeled with words describing the linking relationship.

12. Exercises / Learning activities

• Describe one or more existing DL / DLS in terms of the concepts and constructs of the Digital Libraries Reference Model

• The main actor of this scenario is a digital librarian called D. D has been given the task of designing an information discovery service that integrates 3 different services: Wikipedia, Amazon, and Europeana. The objective of the service is to offer to its users, for a given topic:

  an account of the topic (taken from Wikipedia)
  a list of resources in Europe related to the topic (Europeana)
  a list of the available publications on the topic (Amazon)
  the links between these 3.

• In addition, the digital librarian D should design a mechanism that integrates the three different user models supported by the respective systems:

  • Wikipedia has a user model capturing username and password of the user, as well as his real name. Other features include the names of the groups that the user belongs to and the edits the user has made on the wiki. Optional features include email and gender. When the user is registered he may have access to advanced preferences and editing options. These include the specification of the language in which the site interface will be displayed, the appearance of date and time, as well as user’s time zone.
  • Europeana has a user model that captures username, password, and email address. Additional characteristics include country, language settings, IP address, the date and time the user accessed the website, and the pages the user requested for viewing.
  • Amazon has a user model that includes name, email address, and password. Optional is user’s birth date. Additional characteristics include payment options, shipping address, and a list of past orders.

Questions:
Based on the above information, define the functionality offered by each of the specified DLs in the context of this scenario in terms of the DL Reference Model.

Specify the user models supported by the respective DLs in terms of the DL Reference Model.

13. Evaluation of learning outcomes

In their answers to the exercises, students demonstrate an understanding of:

- The main concepts of the DLRM
- The existing roles in the DL universe

They will further elaborate on functionality and user related issues

14. Glossary

N/A

15. Useful links

1. http://www.dlorg.eu/ - The website of the DL.org project
4. http://www.loa-cnr.it/DOLCE.html - The website of the Laboratory for applied Ontology

16. Contributors

a. EU funded project DL.org - Digital Library Interoperability, Best Practices and Modeling Foundations