L/ECA/DISD/AVLIN/11

8 May 2003

UNITED NATIONS

ECONOMIC AND SOCIAL COUNCIL

Original: English

ECONOMIC COMMISSION FOR AFRICA

Workshop on Technical Aspects of Building Digital Libraries and Electronic Information Networks

Addis Ababa, Ethiopia
10 – 11 May 2003

Digital Libraries and Virtual Information Services: Architecture and Models [1]

[1] Prepared by Abraham Azubuike (ECA)

Introduction

This paper explores the a generalized technical framework for a digital library in which very large numbers of objects, comprising all types of material, are accessible over national computer networks. Eight general principles have emerged from intense exchange of views from researchers. These principles form the key issues in the transition to a true digital library from the network services that we have today. Kahn and Wilensky (1995) elaborated these principles.

The general principles are:

- A technical architecture or model of a virtual/digital library must be based on the applicable legal, economic and social environment.
- The architecture must be based on unambiguous concepts and terminology
- The underlying architectural framework must be separate from the content stored in the library
- Object names and identifiers are the basic elements of any system of storage and retrieval system of a digital library
- A digital object is more than a collection of bits
- The digital library that is used is different from the stored object
- Repositories must look after the object they hold
- Users want intellectual works not digital objects

1. Legal, economic and social framework

The Internet developed on a notion of free exchange of among technical and professional communities. The emphasis was on making information available to colleagues and the public, without charge. The digital library as it is evolving today exists within a much larger economic, social and legal framework. With wide range of intellectual and artistic products being available digitally, transactions on the networks have become more of trading transactions. For example, musical works represent the livelihood of composers and musicians. They require that the integrity of their works be preserved. They also require payment, as do recording studios and concert halls. Such work will only be part of the digital library, if the library supports their interests.

The relevant areas of law include copyright, performance, and other intellectual property, libel and obscenity, communications law, privacy, and international law. Each virtual library operates within a legal and social environment. This must affect the architecture and networking model. Fore example, what constitutes obscenity differs from country to country, so what form of access architecture to deploy must vary. This is a basic technical problem dictated by non-technical milieu.

Designs must ensure the preservation of integrity of digital objects as well as the integrity of the society. Therefore, the architecture should establish clear boundaries between the areas of responsibility of the various parties concerned with the information chain.

2. The architecture must be based on unambiguous concepts and terminology

Terminology proves to be a barrier in describing a digital library. The terminology of networked information systems developed free-style just as cyberspace. The first problem is to distinguish between a digital library and a virtual library, which are now being interchangeably. Simple words mean different things to different people. For example, the words "copy" and "publish" have different meanings to computing professionals, publishers, and lawyers. Common English usage is not the same as professional usage, and the versions of English around the world have subtle variations of meaning.

In a network words should be used very carefully and their exact meaning made clear whenever they are used. An example is "content". In the Kahn/Wilensky architecture, items in the digital library are called "digital objects". They are stored in "repositories" and identified by "handles". Information stored in a digital object is called "content", which is divided into "data" and information about the data, known as "properties" or "metadata".

3. The underlying architectural framework must be separate from the content stored in the library

Almost every type of information can be represented in digital form, including text, pictures, musical works, computer programs, databases, models and designs, video programs, and compound works combining many types of information.

The underlying architecture of the digital library should specify those characteristics that apply to all types of material. For example, every object needs to have a name or identifier; the actions of adding objects to the library or deleting them apply to all material; general purpose methods of security can be provided.

This generalized architecture is a base for extensions that can be tailored for various types of information. The extensions typically include specific formats, protocols, and rights management that are appropriate for the type of material. For example, the extensions for digitized movies will be very different from those for video games; texts are usually described by bibliographic terms, such as author and title, which are of little relevance to a computer program; a protocol designed for interaction with a database is unlikely to be useful in manipulating graphic designs.

Separating general functions from those specific to the type of content has other benefits. It encourages different markets to emerge, and allows a legal framework in which storage, transmission and delivery of digital objects is separate from activities to create and manage the intellectual content.

4. Object names and identifiers are the basic elements of any system of storage and retrieval system of a digital library

Names are needed to identify digital objects, to register intellectual property in digital objects, and to record changes of ownership. They are required for citations, for information retrieval, and are used for links between objects. These names must be unique. This requires an administrative system to decide who can assign them and change the objects that they identify. They must last for very long time periods, which excludes the use of an identifier tied to a specific location, such as the name of a computer. Names must persist even if the organization that named an object no longer exists when the object is used. There need to be computer systems to resolve the name rapidly, by providing the location where an object with a given name is stored. The Corporation for National Research Initiatives has implemented a handle system which satisfies these requirements. A "handle" is a unique string used to identify digital objects. The handle is independent of the location where the digital object is stored and can remain valid over very long periods of time. A global handle server provides a definitive resource for legal and archival purposes, with a caching server for fast resolution. The computer system checks that new names are indeed unique, and supports standard user interfaces, such as Mosaic. A local handle servers is being added for increased local control.

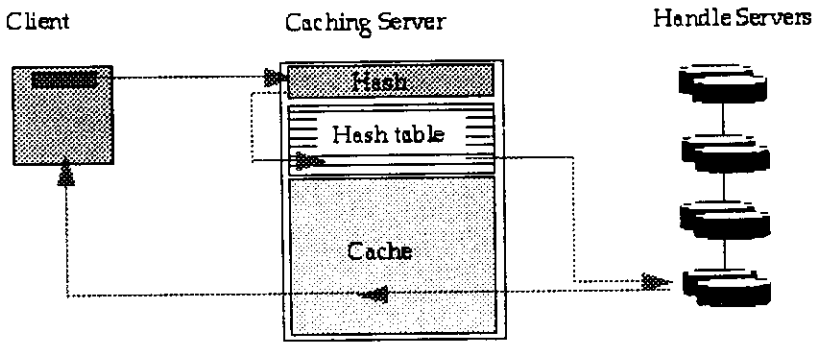

Figure 1. The CNRI handle system

5. A Digital library object is more than a collection of bits

In the digital library, information is stored as "digital objects". A primitive idea of a digital object is that it is just a set of bits, but this idea is too simple. The content of even the most basic digital object has some structure, and information, such as intellectual property rights, must be associated with the digital object. Figure 2 shows that a digital object in a repository has two parts, content and associated data, sometimes called "metadata".

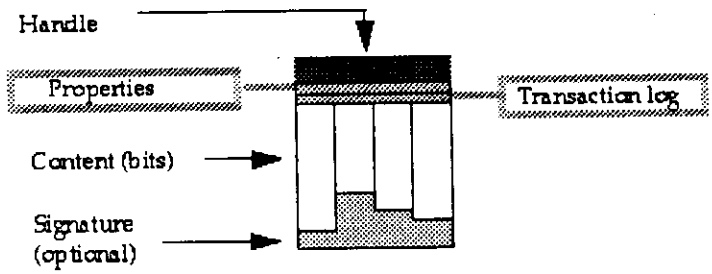

Figure 2. Parts of a digital object

To enable the content to represent useful information, its type must be known. Thus part of the content may be of type: text (perhaps encoded in a mark-up language), while another part may be of type: audio. A single digital object may contain many types of content. It turns out that arbitrarily complex data types can be constructed from a few basic types, notably bit-sequences, handles and other digital objects. By combining these in various combinations, any digital content can be represented.

To manage valuable intellectual property, certain metadata is required. This is shown in the figure. It always includes a unique identifier (the handle). It may also include properties such as rights and access methods. One property states whether a digital object is mutable, in that it may be altered after being placed in a repository. Another is a digital signature or other method of validating that an object has not been changed. Frequently, it is useful to keep a log of all transactions associated with each digital object.

6. The digital library object that is used is different from the stored object

In the digital library, what you store is not what you get. The architecture must distinguish carefully between digital objects as created by an originator, digital objects stored in a repository, and digital objects as disseminated to a user.

The user receives the result of executing a program on the stored object. This may be a simple program, such as a file transfer program, or something very complex. For example, an image is stored in a library as a set of wavelets. To use it, the stored wavelets are used to generate an image with the characteristics requested. This is transmitted over the network to a user's computer, where it can be further processed or displayed.

Some classes of digital objects can be provided to a user in more than one way. For example, the score of a musical work is held in the library. One form of use is to transmit a representation of the score to the user's computer. Alternatively, the user could request the repository to execute a synthesizer program, which would perform the score, and transmit the digitally encoded audio over the network. For some types of object, such as a database or a video game, the use consists of an interaction between the user and the execution of the program.

7. Repositories must look after the information they hold

A repository stores digital objects, both the content and the metadata.

A digital object as stored in a repository may be very different from the digital object that is made available to users' computers. Different repositories will have very different internal organizations, but for each digital object, every repository will have a properties record, which holds attributes of the object, and a transaction log.

Since digital objects contain valuable intellectual property, the stored form of a digital object within the repository includes information that allows for it to be managed within economic and social frameworks. The repository maintains this information, provides basic reference information, and provides security to ensure that only valid operations are carried out on the digital objects.

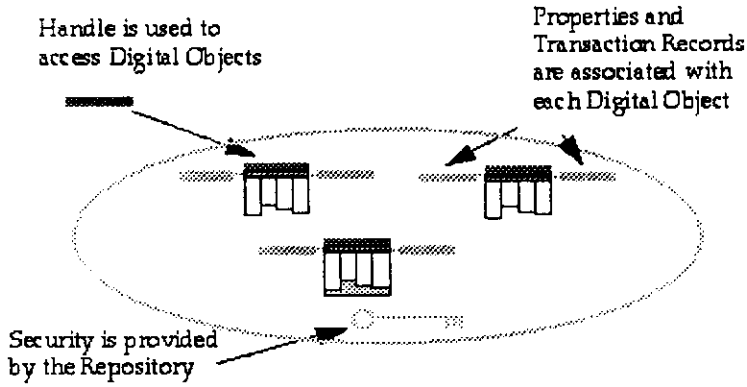

Figure 3. A repository

The internal organization of a repository and the way that digital objects are stored are hidden from the user. A simple protocol is provided for interactions with the repository. This protocol is called the "repository access protocol." The basic commands in this protocol are those to access a digital object and its metadata, and the service request to disseminate a digital object. In addition there are commands to add and delete digital objects.

8. Users want intellectual works, not digital objects

Digital objects are the basic building blocks of the digital library, but users of the library usually want to refer to items at a higher level of abstraction. Common English terms, such as "report", "computer program", or "musical work", often refer to many digital objects that can be grouped together. The individual objects may have different formats, minor differences of content, different usage restrictions, and so on, but certain users are willing to consider them as equivalent.

Which digital objects should be grouped together can not be specified in a few dogmatic rules. The decision depends upon the context, the specific objects, their type of content and sometimes the actual content. The underlying architecture has to support two main needs. It must provide methods for grouping digital library objects and must provide means for retrieval.

The Kahn/Wilensky architecture supports these higher level ideas in several ways. One is to have a digital object containing several digital objects. Thus several formats of a text might be assemble into a single digital object. Another approach is to have these variants stored as separate digital objects, each with its own handle. These handles are contained in a digital object, known as a "meta-object", which acts like a catalog record. It contains a list of the variants with their handles and information about the differences amongst them.

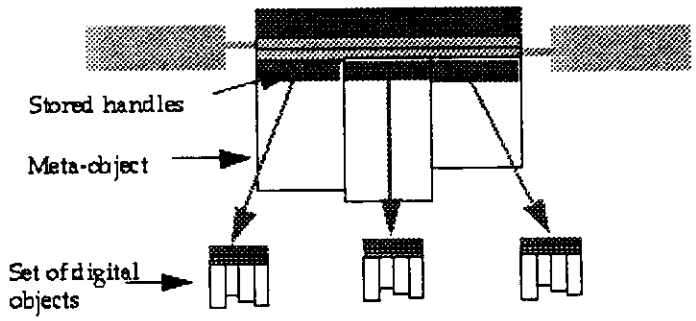

Figure 4. A digital object used as a catalog record

Acknowledgement

This paper is an adaptation of *Key Concepts in the Architecture of the Digital Library* By William Y. Arms, *D-Lib Magazine*, July 1995.
Available from: http://www.dlib.org/dlib/July95/07arms.html. Accessed 2 May 2003

Reference

hdl:cnri.dlib/tn95-01 Kahn, Robert and Wilensky, Robert. "A framework for distributed digital object services". May, 1995. (http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html.)