
NSF/JISC Repositories Workshop
Babak Hamidzadeh, Library of Congress
April 8, 2007

1. Objective

We need to build digital repositories that are capable of preserving and making available a multitude of content types in large size that are received from heterogeneous sources.

The main tasks performed by a digital repository are thus:

Transfer

- The ability to accept a multitude of digital content types in different formats from diverse sources.
- The ability to inspect and analyze the transferred materials.
- The ability to verify the integrity, safety, and authenticity of transferred material.

Appraisal

- The ability to select content, from an available set, for acquisition into library collections.
- The ability to select content by examining individual items or their aggregations.
- The ability to select content by examining descriptions of individual items or their aggregations.
- The ability to do the selection at or within different stages of the content lifecycle.

Preservation

- The ability to store large amounts of digital material over long periods of time.
- The ability to protect digital material from content loss or alteration due to media degradation, technology failure, human error, and natural disaster.
- The ability to migrate digital material across technologies and content types when necessary.
- The ability to model and organize digital material.
- The ability to provide tools for Library staff to curate digital materials, including versioning, meta-data management, content annotation and others.

Access

- The ability to make digital material available to designated users.
- The ability to search for digital materials within collections and across collections.
- The ability to restrict access to digital materials according to business rules and rights laws.

2. Business Case

Costs associated with managing large digital materials and their growth are one of the primary risk factors in information management. The primary cost factors are:

1. Labor costs associated with managing large collections of information in different parts of their lifecycle. This cost is a factor even if we have effective processes and workflows that utilize labor and skills well.
2. Ineffective or flawed processes, workflows and technologies.

3. Approach

The sheer digital content size and the growth rate of this content, the diversity and number of types, formats and sources of the content, and the risk associated with the loss or inappropriate dissemination of the content, dictate that as an initial guiding principle we perform basic but essential functions, efficiently and reliably.

As stated in earlier sections, the basic functions to concentrate on are content transfer, appraisal and preservation, and content dissemination and access over the long term. If we are not able to reliably receive large incoming content, and if we are not able to maintain the content so it remains accessible, useable and understandable in the long term, we will either lose information shortly after its production, or we will have an unmanageable backlog of information that in itself will lead to its loss, or we will have potentially stored content that in few years will be strings of meaningless and useless bits.

To meet the basic objectives of a digital repository, enabling technologies are needed that possess the following characteristics.

- Be easy to operate: Since many of the above functions will be performed by librarians and curators, since the technical products supporting these functions will need to be operated and maintained by technicians and operators, and since the digital materials will ultimately have to be accessible to audiences with potentially limited technical capabilities, the digital library systems that we develop will have to be easy to operate, maintain and to use.
- Enable interoperability with other systems: Any system that we develop will have to interface and interoperate with other existing or future systems. Therefore, it is important that our systems provide clear and easy-to-use interfaces with other systems. Our systems or their parts must be easy to integrate with whole or parts of other digital library systems as well.
- Enable representation of rich interrelationships between digital objects: Maintaining the integrity and accuracy of complex digital objects requires that the interrelationships between the components of complex digital objects be represented and maintained in a machine-understandable way. Representing such

relationships will also provide important context that will be essential for understanding individual components of digital objects as well.

- Be flexible: Since different collections and organizations that manage them will require varying degrees of technical and procedural support, the digital library systems will have to be easily adaptable (either through reconfigurability or by requiring minimal re-engineering and coding) to suit the needs of collections and their managing organizations.
- Unify digital content sets: Enabling the search within and across collections with heterogeneous intellectual content, content types and formats will require a degree of unification and integration in information representation. Unification practices in information representation and management processes can also facilitate preservation in some cases. The aspects of the system to be unified and the degree to which they should be unified will depend on the detailed requirements of each system.