Directions for Deliverables of EDIT's First Reporting period (covering Month 0 – March 2006 to Month 11 – February 2007)

The deliverables are often written reports but can also take another form, for example the completion of a prototype, etc. In such cases the deliverable should nevertheless also be documented in a written record of the achievement of the deliverable in addition to being listed as an achieved deliverable in the Periodic Activity Report, including any available supporting material (e.g. photos of the prototype, the report of the conference….).

Any delay in the submission of a deliverable must be reported in the Periodic activity report, in the section **"Section 2 - Workpackage progress of the period",** where both the due date and the actual submission date (or the foreseen date, if the deliverable is not yet submitted) are reported.

Please note that the following front page is a standard provided by the EC, all requested information on this page must be filled in.

Max. 2 pages (front page excluded) per deliverable in "Garamond" 12 points. As far as possible please do not change the lay out of the standard front page.

Project no. 018340

Project acronym: EDIT

Project title: Toward the European Distributed Institute of Taxonomy

Instrument: Network of Excellence

Thematic Priority: Sub-Priority 1.1.6.3: "Global Change and Ecosystems"

WP 5.3.1 Existing Digital Library activity, principles and standards

Due date of deliverable: Month 12
Actual submission date: Month 12

Start date of project: 01/03/2006 Duration: 5 years

Organisation name of lead contractor for this deliverable: 10 NHML

Revision [draft, 1]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level ("X" in the relevant box)		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Abstract

This document provides an overview of existing Digital Library activity, taking into account established principles and standards in the field. It examines practices and standards related to the accompanying software specification and other deliverables for work package 5.3. of the EDIT project. The report starts by examining the definition of the term 'Digital Library' in the context of the EDIT project, noting that in this scenario, it relates specifically to a systems or set of software utilities that support the management and delivery of electronic literature and other digital objects of use to taxonomists. It examines key functions to fulfil this, and takes an overview of the types of material that are often held in a digital library. It then examines some existing digital library resources that full-fill a similar function. Appendices cover current defined standards for metadata, image creation, text based object delivery and best practice in collection development policy.

1. Introduction

1.1. Definitions: digital and virtual Libraries

The phrase ‘digital library’ has been used to describe any collection of digital material available on the Internet or in a closed networked environment that has some method of organisation. Anything from a website of categorised links and word documents to a full size institutional repository has been described as a ‘digital library’. To achieve a manageable scope for this report and for the following deliverables in WP 5.3, it is first necessary to reach a workable definition of a ‘digital library’. When attempting to reach such a definition, it is also thus necessary to understand exactly what resources a digital library might hold, specifically, the digital objects it contains or provides access to.

Cornell University defines a digital object as “an item as stored in a digital library, consisting of data, metadata, and an identifier” . To further qualify this, a digital representation of a physical book could include data (image files and / or text files comprising the book), metadata (a bibliographic record of the book and structural metadata, describing the links between the pages) and a unique identifier for the object; i.e. a Digital Object Identifier .

The phrase ‘Digital Library’ is often used to refer to systems that organise and make available digital objects held by one or a group of institutions. This concept has been codified in the OAIS reference model , an international standard for the operation of a digital archive.

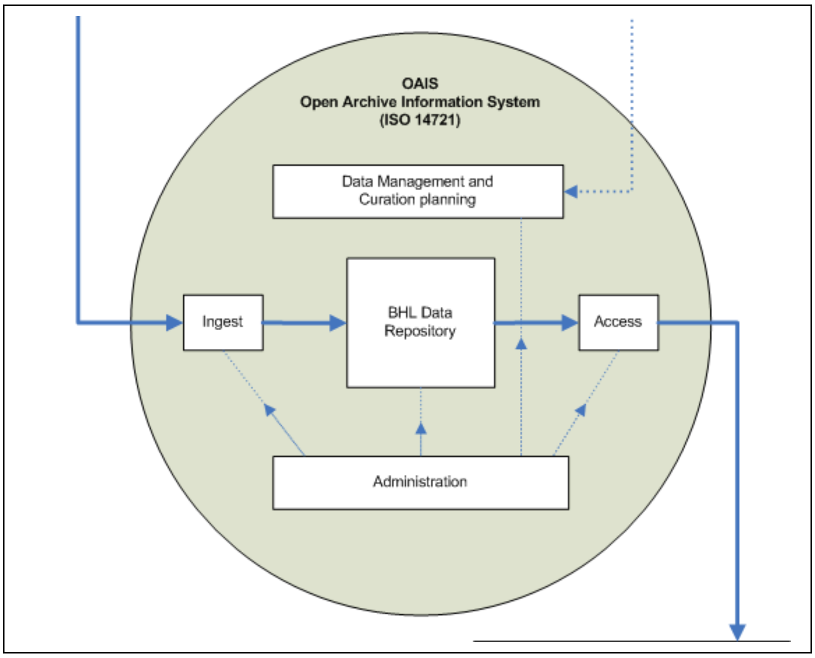
OAIS
Open Archive Information System
(ISO 14721)
Data Management and Curation planning
Ingest
BHL Data Repository
Access
Administration

Figure 1. Conceptual overview of the OAIS standard as a model for digital library operation

This concept differs from the term 'virtual library', often used to denote disparate resources held by separate bodies or institutions that share a common theme or subject area . In the case of the EDIT project, the latter term is more readily applicable to the key deliverable software in work package 5.3, the proposed European Virtual Library of Taxonomic Literature (ViTAL).

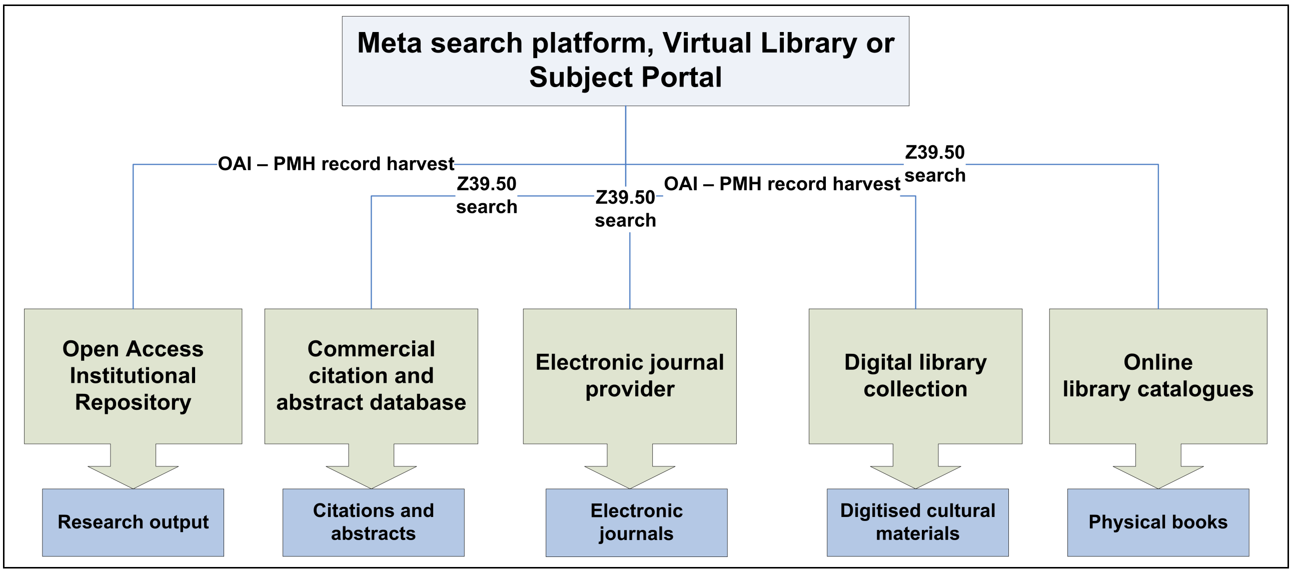
Meta search platform, Virtual Library or Subject Portal
OAI – PMH record harvest
Z39.50 search
Z39.50 search
OAI – PMH record harvest
Z39.50 search
Open Access Institutional Repository
Commercial citation and abstract database
Electronic journal provider
Digital library collection
Online library catalogues
Research output
Citations and abstracts
Electronic journals
Digitised cultural materials
Physical books

Figure 2. Conceptual overview of the 'virtual library' software model

Work package 5.3 will not primarily be involved in the creation and management of new digital objects, but instead focusing upon providing new means of access to resources as created and managed by other entities. These resources would be made accessible to the taxonomist in an organised fashion, based upon the requirements and work-flows defined in previous EDIT work packages.

Following this split in definition between the ‘virtual library’ model and the ‘digital library’, software platforms can then be roughly divided between the two models, into virtual library or portal software and fully integrated digital object creation and management systems. Whilst ViTAL is foremost a ‘virtual library’ project , it is expected that a full specification of software functionality for the ViTAL platform will draw upon areas common to both of the models outlined above. This report will focus upon functionality and content types common to both models.

The rest of this report considers best practice in the following areas:

> Section 2 of this report examines functionality, standards and protocols common to the two models of digital library outlined above.
> Section 3 outlines content held in a typical digital library. Supporting appendices provide greater background information on metadata standards and related practices.
> Section 4 examines some major digital library implementations in operation, also examining future developments in digital libraries.
> Section 5 examines open source and commercial software digital library solutions with two separate case studies.

2. Functionality - key principles of a digital library

2.1 Resource integration

A ‘virtual library’ exists by integrating access to disparate resources held in other digital libraries. There are currently three major record bibliographic standard protocols used to achieve resource integration.

2.1.1. OAI - PMH (Open Access Protocol for Metadata Harvesting)

This is a means of selectively or completely harvesting records from one system to another. It allows for the easy creation of ‘virtual libraries’. Harvests can be run sequentially, ensuring that data-sets held in different locations are readily synchronised. It is used heavily in Open Access research publishing .

2.1.2. Z39.50

This is another means of accessing bibliographic records held in another system. The standard is mainly used in library catalogues. It can be used to query external catalogues with a given type of search .

2.1.3. Open-URL

This is a means of directing users to the specific copies of subscription electronic resources that they are entitled to access. It directly links externally held material to the citations and references in a third party database .

All three methods are employed in single search software platforms used in major research institutions across the world. In order to provide an effective service and tie into as many commercial and open access resources as possible, any viable digital library software solution should ideally make use of these methods to integrate resources held externally, and in the case of digital object management, to allow other virtual library systems access to the resources it may 'manage'.

2.2. Search and retrieval of resources

To effectively search and manage large volumes of information, a single point of search to access all types of material held in multiple repositories is required. This allows a user to simultaneously search multiple sources of digital objects and related reference material at the same time with one search term, and return results from those data sources in one interface. This will allow a user to achieve greater retrieval effectiveness, and to uncover resources that they may otherwise miss. Websites hosting such platforms targeted at a specific subject area are often referred to as 'portals'. Virtual library software platforms that facilitate this are often referred to as 'meta-search' platforms. Such search platforms can offer varying methods of access for different user groups, with varying levels of complexity. Users can also pick and select specific resources to search, rather than search across an entire selection.

In addition to providing a single point of access, some meta search platforms allow for the grouping of resources by subject area. Researchers can even specify which subject areas they would like to search, without having to select specific resources. Advanced search functionality often matches or exceeds that found in modern

Internet search engines, including full Boolean operation, intelligence ranking and de-duplication of results.

2.3. Management and organisation of digital objects

A comprehensive digital library system should be able to effectively manage complex Digital objects that it holds. It should be able to recognise the relationship between the component files and metadata that comprise a digital object.

Portal or ‘virtual library’ solutions should also provides some method of organisation when establishing links to resources managed elsewhere. This could involve the categorisation of ‘third party’ digital objects according to subject area.

2.4. System administration and statistical reporting

Various standard management functions will have to be performed by any digital library system, including recording of statistical data. Collections principle six of the *NISO Framework of guidance for building good collections* states that ‘a good collection has mechanisms to supply usage data and other data that allowed standardised measures of usefulness to be recorded’ . It goes on to examine methods of reporting from these statistics, and their importance in developing the collection and justifying expenditure on future projects. A key management function involves the checking and validation of all data entered in a system. Auditing of all data entered into a system is also specified as a vital procedure for an OAIS compliant repository . Recording, monitoring and reporting of usage data is then, a vital function in developing and refining any digital library service.

2.5. Access control

Any virtual library solution integrating with resources available via subscription based providers should have a means of access control and authentication, preferably through a single login. The system will be able to know which resources a user can access at anyone time, and provide seamless access. It should also be able to distinguish between different levels of access. For instance some commercial full text services allow all users to search records, but only some to view full text images.

2.6. Resource provenance, access restrictions and copyright management

When integrating resources held by other digital library, it is vital to establish the provenance of material. Ownership, publication history and copyright status of a digital object are all of interest to the researcher. Developments in intellectual property law, such as Creative Commons and Scientific Commons Licensing will lead to increasing amounts of material being made available 'free of charge', and with scope to re-use in certain ways. Controlling these varying levels of access will require greater degrees of complexity for the systems managing and providing access to the resources.

3. Digital library content

Web based literature and other bibliographic resources likely to be of use to the taxonomic researcher fall into two distinct categories; those that are freely available and those that are subscription based. Each category can include both full text on-line resources and citations and abstracts that relate to printed material. A comprehensive digital library service will make use of both subscription-based and freely available material.

This rest of section examines common types of content that comprise the digital objects that are freely and commercially available to researchers via digital libraries.

3.1. Metadata

Metadata is described as "*Data about other data, commonly divided into descriptive metadata such as bibliographic information, structural metadata about formats and structures, and administrative metadata, which is used to manage information"* . When applied in the context of a digital object, metadata can indeed be separated the following main areas of focus:

Descriptive - Data describing the original object from which the digital one was derived.
Technical - Data regarding the creation of the digital object and information on the file formats used.
Structural – Information on the relationship between the component files in a digital object.
Preservation - Data regarding the preservation and usage of the digital object as a long-term resource. Preservation metadata also holds information on all of the changes made to the data.
Administrative – Data relating to the creation of the metadata itself. This also holds rights and usage data regarding the object. Note that much of this can overlap with the requirements of preservation metadata.

Most digital objects have some type of metadata that matches the above standards, often in various formats. Appendix one covers current core metadata standards in digital libraries.

3.2. Images

Digital objects are often comprised of single or multiple digital images. A digital library system should be able to provide access to objects comprised of multiple images. Appendix two examines issues related to image creation, display and file choice, whilst appendix one covers related metadata issues. Images are common content in digital libraries. Sites such as NYPL Digital from the New York Public Library offer scanned digital versions of material alongside text based material .

3.3. Text

Several file types exist for storing and delivering text based resources. Appendix three covers current standards in some detail.

3.4. Multimedia content

Digital libraries increasingly hold or provide access to multimedia content alongside text and still images. Modern compression techniques and increases in bandwidth have lead to a large increase in the amount multimedia material being held in research and cultural heritage institutions. Full digital object management systems often require complex tools to effectively manage and disseminate multimedia content. Virtual library style meta-search platforms should also be able to denote the type and nature of multi-media content, and provide information on the type of software required to access them.

3.5. Print material catalogues

Digital libraries often co-exist alongside electronic records of print collections held by research institutions. To achieve seamless integration of virtual and physical resources, some digital library services offer access to print catalogues alongside search and browse access to digital library objects.

3.6. Abstract and citation services

Abstract and citation services that link to printed or electronic versions of journal articles are a core resource for any researcher. As print abstract services have been increasingly superseded by electronic databases, it becomes possible to integrate searching and resource discovery of such resources alongside searching and discovery of printed material and complete digital objects. Many meta-search 'virtual library' applications access resources in this way.

The uptake of freely available citation services such as Google Scholar and Windows Live has demonstrated the popularity of this type of resource amongst international academic communities. Integration with actual electronic resources via the OpenURL link resolver technology is a key enhancement for the provision of on-line abstracts and bibliographic citations.

3.7. Content selection

As well as dealing learning to deal with a range of content types, a key issue in digital library . Appendix four examines best practice in digital library collection development, as a precursor to a complete collection development policy document for WP 5.3..

4. Major Digital Library Activity

This section examine several international digital library projects as cases studies.

4.1. DELOS / BRICKS – infrastructure development

An EU funded infrastructure project, The DELOS network of excellence in Digital Libraries aims to co-ordinate major developments in future Digital Library infrastructure across the Union. Its website states that:

> *"Future Digital Libraries should enable any citizen to access human knowledge any time and anywhere, in a friendly, multi-modal, efficient, and effective way. A*

> *core requirement for such Digital Libraries is a common infrastructure which is highly scalable, customizable and adaptive. Ideally, the infrastructure combines concepts and techniques from peer-to-peer data management, grid computing middleware, and service-oriented architectures ”.*

BRICKS (Building Resources for Integrated Cultural Knowledge Services) is a related EU driven attempt to create software that will allow museums and libraries to integrate digital collections into a shared library. It is a practical exercise with limited funding building upon many of the ideas forwarded by the DELOS organisation and is a part of several DELOS work-packages. It aims for a distributed, de-centralised network of nodes that communicate on a peer-to-peer basis, passing ‘requests’ along one node at a time. It also aims to create an architecture for a distributed library, along with four working applications (or Pillars) aimed at specific sectors within cultural services. One, the scriptorium, is specifically aimed at digital texts .

4.2. MICHAEL – cultural heritage access development

MICHAEL (Multilingual Inventory of Cultural Heritage in Europe) is a key driver in European digital library developments for the cultural heritage sector. It aims to provide comprehensive access to cultural heritage materials in France, Italy and the United Kingdom. MICHAEL is an example of a highly developed portal or virtual library, in that the resources it links to resources held elsewhere, often on the website of a cultural heritage institution.

MICHAEL adds value by its categorisation of resources by subject area and intended audience type. It achieves this through additional descriptive metadata. Vitally for a European project, it also includes spatial coverage and language type. Whilst its business model, subject matter and target audience differ from that of EDIT, MICHAEL stands as an example of how a ‘virtual library’ can tie together disparate resources to create a comprehensive cohesive resources for varied audiences, creating new access points for existing material and drawing in new users.

4.3. Biodiversity Heritage Library – mass digitisation for the life sciences

The Biodiversity Heritage Library (BHL) is an ambitious ‘bulk digitisation’ project aiming to provide a single, comprehensive resource on bio-diversity based upon digitised collections from major research universities across the world .

The project is aiming to digitise the greater proportion of public domain English language material in the life sciences. Out of copyright material in this field is still of great potential use to the taxonomic researcher. It would become more useful if they can access it via the web at the point of research. The Natural History Museum in London and Kew Gardens are the two main European partners, alongside and American institutions including the Smithsonian and the American Natural History Museum. A key partner in scanning and data warehousing is the Internet Archive , and the project forms part of the Open Content Alliance which includes Yahoo and Microsoft as members.

4.4. Open Access Research publication movement – a new publishing model

One of the biggest growth area's in current digital library activity is that of Open Access research publishing. Open access publishing fulfils several functions for an academic and research institutions, including greater exposure of its research content and greater control and management of intellectual property. A full definition and description of open access publishing can be found at the E-prints website .

Digital repository systems such as D-space and E-prints from the basis for most institutional repositories, and can also be used as effective digital object management systems. Such systems will be a key driver in research publication in the future. Virtual library portal solutions must ensure that they can provide access to objects hosted in these systems. The OAI-PMH protocol is a key technology in achieving this.

4.5. Future developments

This section outlines some recent developments in digital library technologies.

4.5.1. Distributed storage and management

As the first integrated digital library systems are put into place by institutions, the next phase of development appears to be focused at the national and international collaborative level. An early indicator of the scope of large scale digital library implementations is the LOCKSS (Lots Of Copies Keeps Stuff Safe) project. This is a development in the storage or electronic journals and other digital material, originating in America and now in use worldwide. It provides open source, peer to peer sharing software that allows resources previously only available from one location to be cached and managed across multiple personal computers.

4.5.2. Semantic web development

Advances in metadata standards and the evolution of a semantic sub-structure for the web and related electronic resources will also have a major impact upon digital library systems. In the semantic environment, each identifiable resource on the web and each of its major descriptive facets (i.e. subject, creator, originating organisation, place of origin etc.) will be represented by a unique identifier. The semantic structure between these identifiers will be represented by RDF, the Resource Description Framework language .

4.5.3. GUID developments

The major development in terms of semantic interoperability for the Life Sciences is the Life Science Identifiers (LSID's) programme, aiming to provide biological data and entities over numerous resources with a common method of identification. The method of creating and resolving these identifiers is under heavy development.

"The LSID concept introduces a straightforward approach to naming and identifying data resources stored in multiple, distributed data stores in a manner that overcomes the limitations of naming schemes in use today. Almost every public, internal, or department-level data store today has its own way of naming individual data resources, making integration between different data sources a tedious, never-ending chore for informatics developers and researchers. By defining a simple, common way to identify and access biologically significant data, whether that data is stored in files, relational databases, in applications, or in internal or public data sources, LSID provides a naming standard underpinning for wide-area science and interoperability."

5. Digital library software platforms

This section briefly examines Digital Library applications for digital object delivery and 'virtual library' meta-searching applications and general digital object management and delivery systems. It examines both open source and commercial solutions.

5.1. Open source software

Open source applications exist to fulfil core digital library functional requirements. This section examines some meta-search platforms and technologies, and one integrated object management solution.

5.1.1. Meta-search platforms

Several open source meta-search platforms exist. Applications such as Helios allow for combined searching through commercial search engines such as Google and Yahoo. Whilst powerful and flexible, they lack the ability to access subscription based bibliographic data services that are 'hidden' from such search engines. They also lack the access management facilities, and ability to query bibliographic data sources via Z39.50 and OAI-PMH.

Separate open source Z39.50 clients also exist, such as Yaz and have readily available programming 'code libraries' . OAI-PMH libraries also exist, as do basic open source OpenURL link resolvers libraries . Whilst open source options exist to full-fill most of the basic requirements for a meta-search platform, no single solution is readily available 'off the shelf'. Substantial development would be required in order to provide an effective integrated and useful service for the taxonomic researcher.

5.1.2. Greenstone – comprehensive digital object management

A multi-platform open source digital library system designed using open standards to manage digital materials and make them available via the web, or via CD-ROM. Greenstone is developed as part of the New Zealand Digital Library project, which aims to "*develop the underlying technology for Digital Libraries and make it available publicly so that others can use it to create their own collections*" .

Two versions are in active development, Greenstone 2 which is stable and based around an XML container format for digital objects and designed to work on as many platforms as possible, and Greenstone 3, a fully-featured multi-layered modular system designed to work using current technologies such as XML, XSLT SOAP and Java. It also makes use of a relational database to store and index metadata. It is designed to be scalable from a desktop computer to an entire corporate library. Collections created in Greenstone 2 are fully transferable to Greenstone 3.

Key features

Greenstone collections can be harvested via OAI-PMH and in turn, the system can act as an OAI harvester.

Greenstone is designed to work 'out of the box' on as many platforms as possible.

It can handle resources in many formats, including common multimedia formats.

An internal XML metadata format holds information regarding the structural relationships between files that comprise an object.

It supports qualified and unqualified Dublin Core as an internal metadata standard, and can import and export a variety of formats.

It also has strong multi-lingual support.

Both versions of the system are still in active development with a strong user base. Many institutional projects including the Chopin Early Editions Online project have used Greenstone for the management and delivery of digital material.

5.2. Commercial solution

This section examines two commercial solutions from an single supplier, the Israeli company Ex-Libris . Whilst there is an increased initial cost, benefits of using a commercial solution include dedicated support and development from an established company with experience in installing and configuring such systems to meet different user needs.

5.2.1. Meta-lib – portal software

Meta-lib is an example of modern commercial bibliographic portal software. It is marketed as a means to integrate hybrid and varied resources. Detailed information on a Metalib implementation can be found in a paper from Loughborough University .

Key features

The product makes use of all three major linking technologies outlined in section 2 of this report.

OpenURL compatibility is achieved through integration of the SFX link resolver. Ex-Libris were key drivers in the development of OpenURL.

It is also designed to work with all major subscription based content providers.

OAI-PMH compliance allows for the harvesting and indexing of open access publishing repositories.

User verification, access control saving of personal preferences, email alerts and other features are available to add-value to meta-search functionality.

6. Conclusions

To full-fill a virtual library role and create a portal with meta-search functionality, EDIT requires portal software to organise and maintain material held and managed elsewhere. A full digital object creation and management system such as Greenstone will not be necessary for the scope of the project, instead, a portal style solution will suffice in uniting disparate sets of digital objects and electronic bibliographic resources into a central organised platform of taxonomic literature.

Based upon some of the findings of this report and examples of digital library implementations, successful deployment and usage of the software deliverable will likely depend upon the following factors:

a) Full specification of user needs to inform system choice
b) Choice of a sustainable system
c) Complete customisation of system based around user's requirements
d) Defined collection development policy, that outlines the resources to be made available via the platform
e) Adequate testing period
f) Involvement of a defined user base throughout the project

Appendix 1 - Metadata standards

The following subsections will examine the major accepted content and encoding standards for each of these key areas of metadata.

A1.1. Descriptive metadata

Cultural institutions often have set guidelines over the content of descriptive metadata. Where possible, in house procedures using accepted standards such as AACR should be followed, with the application of chosen name authority files for particular fields, such as individual names . Where possible, current institutional policies on catalogue record creation should be followed for the content of descriptive metadata.

MARC

The MARC21 record standard has now gained international acceptance as a means of delivering AACR based metadata content. Its XML incarnation, MARC21-XML can be seen as a preferred means of exporting and importing MARC records between systems. Support for MARC and its XML counterpart should be seen as a necessity in any digital / virtual library implementation.

MODS

MODS (Metadata Object Description System) is a subset of MARC, designed specifically for the description of digital objects and expressed through XML. Unlike MARC, it is based upon named elements in English rather than numerical codes. It includes a hierarchical structure of parent and child elements. Oxford University makes use of MODS as the descriptive metadata element of their Digital Library .

Dublin core

The unqualified Dublin Core metadata standard has also been commonly accepted as a base level framework for the sharing of descriptions of digital objects. Any organisation that wishes to share its metadata should provide mappings from its preferred standards to Dublin Core as a basic level of export for record sharing and harvesting. This practice has been ratified under JISC guidelines in the UK .

A1.2. Technical metadata

Technical metadata in the context of a digital library relates specifically to the creation of digital images or digital text from original works, although data can be held on other file types. It can hold information on the means and process of digital capture, the format, size and quality of the digital image.

NISO MIX

Technical metadata standards for images is still very much development. Many projects are using proprietary standards and schemas to hold this information. The Library of Congress is currently developing NISO Metadata for Images in XML (NISO MIX), an XML based schema to encode data based around NISO Z39.87-2002 , an international technical metadata standard for still images.

MIX breaks down the NISO standard into four main areas:

I) Basic image parameters
II) The activity surrounding the creation of the image
III) The change history of the image itself
IV) A means to record changes to the quality of the image after any alteration or migration

Some of these areas, such as change history, overlap with other metadata standards such as PREMIS.

Technical metadata creation

There are growing calls for technical metadata to be automatically generated, either by software that extracts information from a file or at the point of file creation. The National Library of New Zealand Metadata Extract Tool is one developed method of doing this, and has been adapted to work with a number of file formats from Word documents to TIFF files. Metadata on the techniques, equipment and settings used to create an image could also be provided by the equipment itself at the point of capture. Whilst many devices create such data, the lack of industry standards in preservation and technical metadata has limited its effectiveness. JPEG image files often contain metadata in the EXIF format, although much of this is lost upon the point of creation. The RLG Automatic Exposure project engaged in a dialogue with equipment manufacturers and produced a number of papers and tools to facilitate a move to such standards. The JHOVE file validation project also automatically produces technical metadata to NISO standards .

A1.3. Metadata containers and structural metadata

METS

METS (Metadata Encoding and Transmission Standard) is designed as a container with separate sections for descriptive, administrative (including preservation, technical, rights and provenance based data), structural and behavioral metadata relating to a digital object. It is expressed in XML and is capable of holding any XML based metadata system inside its sections.

It has already been adopted by a large number of Digital Library providers .

Whilst it is primarily used to package up metadata content for transmission between systems for the ingestion, sharing and archiving of digital objects, METS also contains its own elements for expressing the structural relationship of component files within a digital object. If such data is held in any other way, it should be constructed in such a fashion as to map to the METS structural section where METS is the proffered method for transmission.

A METS record maintains information on the files themselves associated with an object and location. Much of this also constitutes technical and administrative metadata surrounding the object.

Despite its ability to hold complex structural data, The METS format is not a tool in its own right, but a file format only. Any functionality derived from this format will have to be programmed, as described below;

“METS’s comprehensiveness and the flexibility designed in its structure make it an excellent choice for a framework or container for the objects and metadata in a preservation system. METS is not a tool, however. An instance of METS is an XML document. To be able to work with

METS as a container for Ingest, you need a text editor, an XML editor, or ideally, a forms-based user interface built and customized to your collections and to your working environment. Batch processing will require some customized programming to integrate your metadata into the METS structure. Using it as the container for an Archival Information Package will also require programming work."

Many current systems on the market make use of METS to some degree, often as an import or export format. The Fedora digital resource management architecture also has its own XML based format to describe the relationship between components of a Digital object. Greenstone 2 Digital Library system also makes use of its own internal XML format .

A1.4. Preservation, Rights and other administrative metadata

PREMIS

PREMIS (Preservation Metadata Implementation Strategies) is a relatively new development from a working group of experts in the field of Digital Preservation. According to the groups' website:

(Preservation metadata) is information that supports and documents the digital preservation process.

(Preservation metadata) *addresses provenance (who has had custody/ownership of the digital object?); authenticity (is the digital object what it purports to be?); preservation activity (what has been done to preserve the digital object?); technical environments (what is needed to render and use the digital object?); rights management (what intellectual property rights must be observed?).*

(Preservation metadata) is an essential component of most digital preservation strategies.

PREMIS aims to address to the lack of standards in both content and encoding of such data. It has four separate entities used in object description, the object itself, agents using the object, rights surrounding the object and events affecting the object.

PREMIS can be expressed in XML and has separate schemas for each of the above entitles, as well as an overarching container schema. Whilst PREMIS is a new format in the emerging field of digital preservation, its complete approach and modular nature give it some leeway in its application. Some of the semantic units within the four PREMIS entities appear to overlap with existing metadata schemas in other areas, (i.e. some elements of the object entity overlap with the METS 'filesec' file section.) Monitoring of practical applications of PREMIS as the standard develops is recommended to see how such as complex system is applied in working conditions.

The OAIS specification documentation goes into extensive detail upon metadata requirements for archival information packages that are to be stored in any OAIS compliant archive or repository. It defines a need for Preservation Description Information divided into '*four types of preserving information called Provenance, Context, Reference, and Fixit*y'. Whilst such data may not be required in the dissemination of material form the archive, it is required throughout the archival process itself.

The Dublin Core administrative component is another administrative metadata standard, designed to ensure 'interoperability between systems with content metadata.' It may contain information regarding the creation, transmission and reproduction of metadata. Whilst PREMIS relates to the administration of the whole object, much of the administrative component relates to the handling of the data itself. Some of its functionality is overlapped by other metadata standards, for example, the header section in an OAI-PMH file transfer.

Another system designed to establish the provenance of a digital object, (i.e. its format and adherence to that format's standards and important characteristics) is the JSTOR/Harvard Object Validation Environment (JHOVE) . It validates digital objects against the exact byte specifications of the file format. It also reports automatically generates technical metadata by reporting on the characteristics of a digital object in compliance with NISO Technical metadata for images standards. It exists as a JAVA object for integration into a Digital Library environment, and can be used in the ingestion and validation of Digital objects. A discussion on the integration of JHOVE into a Digital repository system such as Dspace is currently underway at the Dspace Wikki . Ensuring the validity of any digital object is vital for its long-term preservation.

A1.5. Extensible mark-up language (XML) for metadata transfer

The Extensible Mark-up language is now widely regarded as a preferred method for the transfer of electronic data. When using XML based metadata container systems, a preference for the new XML Schema method of verification over the older Document Type Definition should be shown. XML Schema provides full support for the concept of Namespaces, a unique identifier for each type of schema that allows it to be distinguished from others. This allows documents containing data stored in multiple container formats to be kept and transferred together, (as in a METS file). The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) makes use of XML as the means of transferring records between harvesters and repositories.