National Library of Medicine
Digital Repository

Policies and Functional Requirements Specification

Prepared by
the NLM Digital Repository Working Group

Version 1
Submitted January 31, 2007
Revised March 16, 2007

Document control

Revision history

Revision	Date	Author	Reason for change
Version 1	10/30/06	Digital Repository Working Group	Working draft
	12/22/06	Digital Repository Working Group	Final draft
	1/31/2007	Digital Repository Working Group	Final version
	3/16/2007	Digital Repository Working Group	Editorial changes to TEI recommendations in Policies Section 6.2.

Requirements Specification sign-off:

Sheldon Kotzin, Associate Director, Library Operations	March 16, 2007

Acknowledgements

This document uses a Table of Contents that has been adapted from the Volere Requirements Specification Template.

The National Library of Medicine acknowledges that this document uses material from the Volere Requirements Specification Template, copyright © 1995 - 2004 the Atlantic Systems Guild Limited.

Table of Contents

1 Introduction

1.1 The Need for an NLM Digital Repository

The mission of the National Library of Medicine is to acquire, organize, preserve and make accessible the published record of the biomedical sciences and health professions, irrespective of format. In order to fulfill this mandate, it is essential that the Library develop the robust infrastructure needed to manage and preserve a large amount of material in a variety of digital formats.

A number of program areas within Library Operations are in need of such a digital repository to support their existing digital collections and to expand the ability to collect a growing amount of born-digital resources. Dozens of digital collections which require long-term management and preservation have already been created by the History of Medicine Division. Collection development and acquisitions staffs are seeing an increasing availability of born-digital materials, including licensed e-journals and web sites that NLM needs to add to its collection. NLM's preservation program has embraced digitization as a preservation method to replace microfilming. The creation of functional requirements and identification of key policy issues for an NLM Digital Repository are essential next steps to aid in building NLM's collection in the digital environment.

The Library currently has processes in place for the ingest, management, storage, archiving and access of digital material for the following acquisition and ingest streams: electronic journals deposited in PubMedCentral (PMC); digitized back files of PMC journals; NLM Web page archive; and CIT Videocasts. At the current time, these are the only processes that have functionality that addresses the major components of OAIS functionality.

The Library needs to put in place a reliable repository for the preservation of digital content not covered by PMC and the videocast project to ensure ongoing access. Preserving digital material presents a different set of technological challenges from those of preserving analog material. The creation of a comprehensive Digital Repository capable of ingesting digital material, storing and managing that digital material and associated metadata, and ensuring that digital material is accessible for the long term is essential to NLM fulfilling its mission.

The NLM Digital Repository is envisioned as one or more electronic storage systems within which digitized and born-digital objects created or acquired by NLM reside. The repository has the ability to store, preserve and provide access to all types of digital objects. Functionality includes:

- ingest and management of content as well as the descriptive, administrative and structural metadata associated with stored objects;
- preservation of objects in approved formats;
- migration to new formats to insure objects do not become obsolete;
- controls to insure only permitted access to objects.

1.2 The Purpose and Scope of this Document

A Working Group was established in May, 2006 to develop high-level functional specifications for an NLM Digital Repository for NLM collection materials and to identify policy and management issues related to the creation, design and management of the repository.

The scope of the working group's responsibility was limited to digital repository requirements and issues related to the management of collection materials under Library Operations responsibility. This included: the maintenance of the inventory of existing and planned digital projects; development of functional requirements for digital repository needs; identification of metadata and format standards for digital materials (scanned as well as born-digital); and identification of policy issues related to the digital repository. The group considered functionality of software and systems used to handle existing digital collections (e.g., PubMed Central, the Bookshelf, Images of the History of Medicine and other HMD digital collections) in scope.

Digital projects, as well as databases, which manage resources that are not part of Library Operations' collecting responsibility (e.g., the UMLS, Visible Human, image library projects in LHC) are outside the scope of these requirements.

1.3 Use of the OAIS Model to Develop the Requirements

This document provides a high-level statement of the Library's requirements, as known in December 2006. The requirements are based upon the Reference Model for an Open Archival Information System (OAIS)[1] . The requirements are documented at a high level, and the rationale for any addition or modification to the detailed OAIS functions has been noted. The level of detail of the requirements varies across functions in this document and is based on the group's knowledge and understanding of needs at the time. Future iterations of this document may contain additional requirements as they become known.

2 Users of the NLM Digital Repository

The users of the NLM Digital Repository include both end-users, such as the general Public and NIH staff, and NLM staff who will be working with and managing the content and repository system. A crucial feature of a digital repository created and maintained by NLM will be providing the various access levels required by the different types of users. NIH staff may need different access levels to some materials than the general public. NLM staff and system administrators will require access to the NLM Digital Repository in order to ingest, administer, manage, preserve and access objects and their supporting information and structures. This will require multiple levels of access to the materials. In addition, a primary goal of the NLM Repository is to provide access to the material as an active, regularly used archive, as opposed to a dark archive accessible only under certain exceptions.

[1] Consultative Committee for Space Data Systems (2002). "Reference Model for an Open Archival Information System (OAIS)". CCSDS 650.0-R-1 – Blue Book. Available at: http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

2.1 End-User Access

End-users will need to access and use the materials residing in the NLM Repository. Both onsite access and Internet supported off-site access is needed for end-user access to the repository. End-users will be able to search metadata and full text within documents (when available), and view materials in formats readily available to a wide audience such as PDF, JPEG, and XML. However, some end-users may require different access rights to materials. The 2 major end-user groups that have been identified are:

- **General public**, including both United States users and international users. Public access to some materials may be restricted by licensing terms, embargo periods, copyright, etc. This level may be sub-divided into location level access when materials are only available via specific work stations or areas, such as the NLM Reading Room due to subscription or purchase requirements and licenses. There may even be instances where materials are available to U.S. citizens and not international users.

- **NIH staff**, including NLM staff, which need to use the repository materials for research and reference. This level may be sub-divided into location level access when materials are only available via specific work stations or areas, such as the NLM Reading Room due to subscription or purchase requirements and licenses.

2.2 NLM Staff Access

A broad NLM staff level access is needed for NLM staff working with the repository and the objects in it. The 2 identified NLM staff categories are:

- **Metadata Access**, for staff creating, reviewing or editing appropriate metadata about the materials in the repository. This level will allow NLM staff to view materials and add to or edit metadata without changing the object itself. There should be multiple levels of access for viewing and editing of data based on varying operator and metadata classes.

- **Object Maintenance Access**, a more restricted NLM staff level, available to staff working with the object itself: adding new files, checking the quality of digital files, manipulating images, performing format conversions and migrations (e.g., from tiff to pdf, tiff to jpeg2000, etc.), and investigating problems with the system.

2.3 NLM System Administration/Programming Access

The most restrictive NLM staff level providing ultimate rights to the system, required for its management, development, and assigning appropriate rights to users.

The following chart of Operational Scenarios and User Classes (Figure 1) illustrates the various types of user interactions expected with the NLM Digital Repository. Ingest Preparation which is listed in this chart is not a component of the OAIS model but has been included as a necessary function for the NLM Digital Repository. Further information on ingest is listed in Section 7.1.

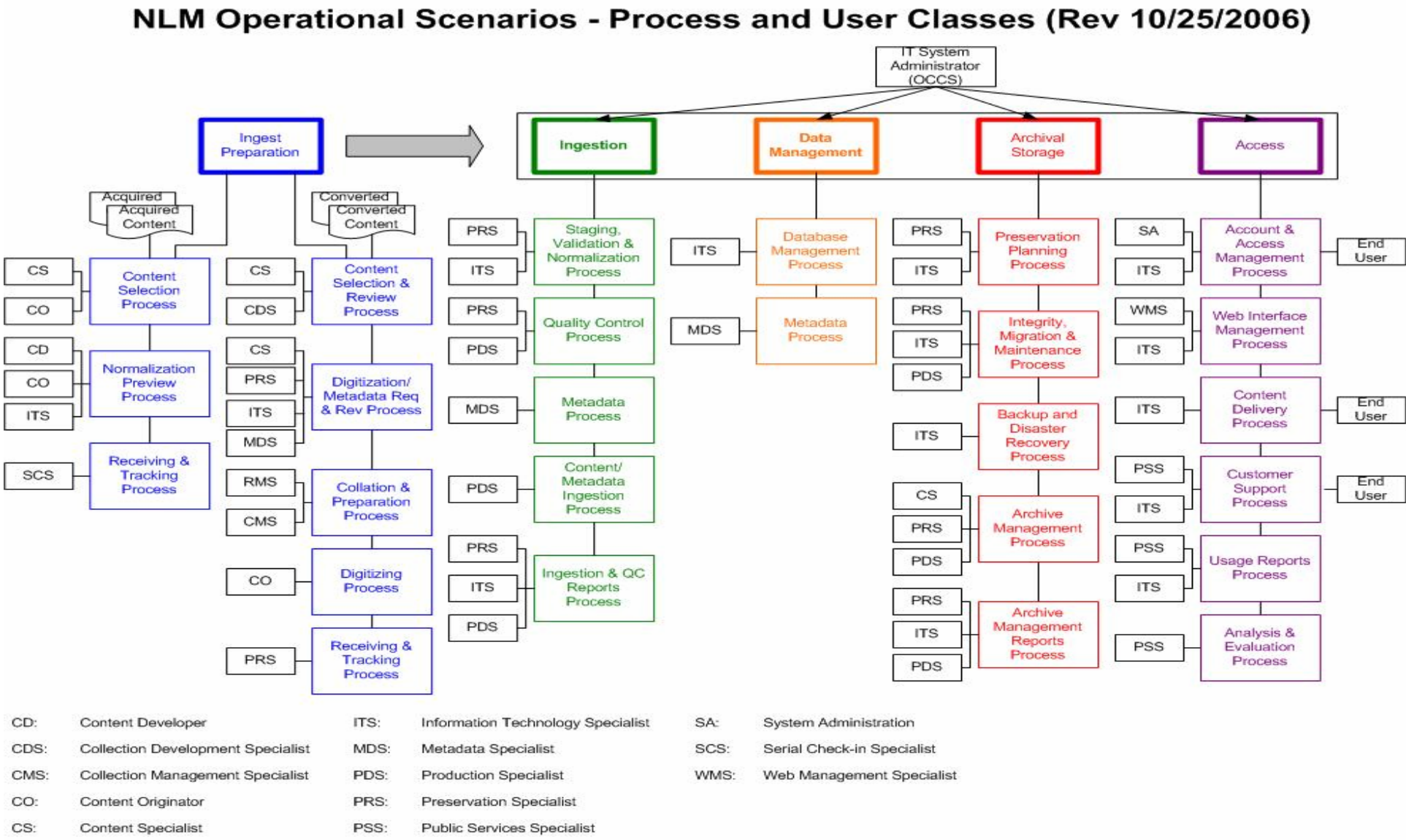

Figure 1 NLM Operational Scenarios – Process and User Classes

3 Project Constraints

3.1 Design Constraints

- Existing digital repositories for selected portions of the NLM Collection (e.g., PMC, Bookshelf, NIH Videocasts.
- Lack of standards for file types (e.g., e-books)
- Materials found in NLM's special collections may require special consideration concerning the type of formats and data files that must be stored within the repository. Important donations of historical materials may include obsolete or unusual formats.
- Bibliographic data will be supplied by NLM's Integrated Library System (ILS). The repository should provide an interface to this data.
- Digital material held in the NLM Digital Repository will be a vital asset for NLM staff and the world. It is therefore important that a suitable set of security controls and procedures is achieved. NLM follows the HHS/NIH security standards.
- The Library will follow the HHS/NIH/OCCS security policies for the NLM Digital Repository.
- The security framework must address functioning with the NIH off-site remote backup facility (NCCS), in both back up and parallel mode.

3.2 Implementation Timeframe Constraints

- Phasing: Evaluation and testing of selected commercial and open source software options is expected to occur prior to selection and implementation of a production repository.
- Budget: Budget costs are unknown until evaluation of software options is completed. Depending on the solution selected, funding may be needed for purchase of software and/or additional programming support.
- Staff resources: Lack of availability of Library Operations systems staff to work on digital repository testing and implementation.

4 Standards, Definitions and References

4.1 Use of Standards

- The NLM Digital Repository must be designed with regard to international best practice for digital archives/repositories.
- As new standards are developed NLM will need to review them for use by the repository.
- The following standards are mentioned in this document:
    - Dublin Core
      (http://dublincore.org/)
    - Encoded Archival Description (EAD)

(http://www.loc.gov/ead/)
- MARC21 and MARCXML
  (http://www.loc.gov/standards/)
- Metadata Encoding and Transmission Standard (METS)
  (http://www.loc.gov/standards/mets/)
- Metadata Object Description Schema (MODS)
  (http://www.loc.gov/standards/mods/)
- NLM DTDs – journal archiving, book, historical book, and NLMCatalog
  (http://dtd.nlm.nih.gov/), (http://www.nlm.nih.gov/databases/dtd/)
- ONIX
  (http://www.editeur.org/onix.html)
- The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
  (http://www.openarchives.org/OAI/openarchivesprotocol.html)
- OpenSearch
  (http://www.opensearch.org/)
- OpenURL
  (http://www.niso.org/standards/standard_detail.cfm?std_id=783)
- PREservation Metadata Implementation Strategies (PREMIS)
  (http://www.loc.gov/standards/premis/)
- Search/Retrieve via URL (SRU) and Search/Retrieve Web Service (SRW)
  (http://www.loc.gov/standards/sru/)
- Text Encoding Initiative (TEI)
  (http://www.tei-c.org/)
- Z39.50
  (http://www.loc.gov/z3950/agency/)

4.2 Definitions

A Glossary of Terms is provided as Appendix C.

4.3 References

The following is a list of references used in developing this document.

Association of Research Libraries. SPEC Kit 292: Institutional Repositories (2006)
http://www.arl.org/pubscat/pr/2006/spec292.html

Coyle, Karen. Rights in the PREMIS Data Model: A Report for the Library of Congress (2006)
http://www.loc.gov/standards/premis/Rights-in-the-PREMIS-Data-Model.pdf

Cornell University Library. Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems (2004)
http://www.library.cornell.edu/iris/tutorial/dpm/index.html

Data Dictionary for Preservation Metadata. Final Report of the PREMIS Working Group. (2005)
http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Electronic Resource Preservation and Access Network. ERPA Guidance: Ingest Strategy. (2004)
http://www.erpanet.org/guidance/docs/ERPANETIngestTool.pdf

McLeod, R. and Wheatley, P. and Ayris, P. Lifecycle information for e-literature: full report from the LIFE project. Research report. LIFE Project, London, UK. (2006)
http://eprints.ucl.ac.uk/archive/00001854/

National Digital Information Infrastructure and Preservation Program (NDIIP). Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program (2002)
http://www.digitalpreservation.gov/about/planning.html

National Library of New Zealand. National Digital Heritage Archive Programme Business Requirements Specification (2005)
http://www.natlib.govt.nz/

NISO Framework Advisory Board. A Framework of Guidance for Building Good Digital Collections, 2nd ed. (2004)
http://www.niso.org/framework/Framework2.pdf

Open Society Institute Guide to Institutional Repository Software, 3rd ed. (2004)
http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf

Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects. OCLC/RLG Working Group on Preservation Metadata (2002)
http://www.oclc.org/research/pmwg/

Preserving the Past to Protect the Future: The Strategic Plan of the National Archives and Records Administration 2006-2016 (2006)
http://www.archives.gov/about/plans-reports/strategic-plan/2007/nara-strategic-plan-2006-2016.pdf

Reference Model for an Open Archival Information System (OAIS) (2002)
http://public.ccsds.org/publications/archive/650x0b1.pdf

Technical Evaluation of Selected Open Source Repository Solutions in New Zealand (2006)
https://eduforge.org/docman/view.php/131/1062/Repository%20Evaluation%20Document.pdf

United States. Government Printing Office. Future Digital System (FDsys) Requirements Document, version 2.0 (2006)
http://www.gpo.gov/projects/pdfs/FDsys_RD_v2.0.pdf

5 Relevant Assumptions

The following assumptions have been identified by the Digital Repository Working Group as relevant to this Requirements Specification.

#	Assumption
1	The NLM Collection Development Policy defines the scope of the NLM Digital Repository.
2	A primary goal of the conversion of content for the repository is the preservation of the content. Priority will be given to preserving the intellectual content. The preservation of the original "look and feel" of the original source will be considered and reviewed on a case by case basis.
3	NLM originals will be kept in the collection after digitization.
4	NLM will not provide or maintain special software required to access materials in the NLM Digital Repository (e.g., for audiovisuals or computer applications).
5	The NLM Digital Repository will contain materials digitized by NLM and born-digital materials created or acquired by NLM.
6	NLM will undertake a distributed custodial model for some digital materials. Not all digital objects will be stored in the NLM Digital Repository, but the metadata for those objects will be stored in an NLM-managed system, such as the Integrated Library System (ILS).
7	At a minimum, the original normalized object, the most recent migrated version and the next-to-the-last migrated version will be kept. Other intermediary versions may be discarded, and in some cases all versions may need to be retained.
8	All digital objects must have an NLM unique internal identifier (UID).
9	The system design and architecture should minimize duplicate data entry of any metadata for objects in the NLM Digital Repository.
10	Master images and descriptive/structural/administrative metadata that meet NLM's specifications are created for all digital content that is produced from NLM analog collections and intended for permanent retention.
11	The NLM Digital Repository will contain objects of varying levels of permanence, including ones that will not be permanently retained. NLM will assign permanence ratings to objects stored in the NLM Digital Repository.
12	The Repository must support the use of multiple languages and non-Roman scripts for ingest, maintenance and access of digital objects and associated metadata.
13	Web harvesting functionality is outside the scope of the Digital Repository requirements. NLM expects to use separate software for web harvesting. Harvested data/web pages will be stored in the repository.
14	The NLM Digital Repository will not be a dark archive. Objects in the repository will be available for use, although access to specific objects may be restricted to on-site use or for a set period of time due to embargo.

15	The Repository system must follow HHS/NIH standards concerning software and systems security.
16	To the extent possible, automation should be used for the extraction of descriptive and technical metadata.

6 Policies

6.1 Existing Policies Related to the Digital Repository

- *Collection Development Manual of the National Library of Medicine*
  The *Collection Development Manual of the National Library of Medicine* (CDM), fourth edition (2004), establishes boundaries for the Library's permanent collection and provides a conceptual and philosophical framework for the selection of biomedical materials. It is intended primarily as a working document for NLM staff and defines the range of subjects to be acquired and the extent of the Library's collecting effort within these subjects. It also addresses selection issues presented by a range of formats and literature types.

  The CDM is based on the Collection Development Policy adopted by the NLM Board of Regents in 1976 and amended in 1983 and 1992. The NLM Board of Regents policy statement that "the National Library of Medicine has the responsibility for acquiring the biomedical literature in any format deemed appropriate to the fulfillment of its mission" provides the underpinning for following the CDM policies for selection of materials for the NLM Digital Repository (http://www.nlm.nih.gov/tsd/acquisitions/cdm/).

- *National Library of Medicine Selection Criteria for Digital Reformatting*
  The *National Library of Medicine Selection Criteria for Digital Reformatting* (http://nlm.nih.gov/psd/pcm/digitizationcriteria.pdf) sets the priorities for the selection of materials from NLM's collection to be digitized both for access and for preservation. The goals of NLM's digitization program are to increase access to its collections and provide wider use of its rich resources in the areas of medicine, health care and public health while preserving rare, valuable and fragile materials.

  Selection of materials for digital reformatting is based on monetary, scholarly and historical value; bibliographic control; frequency of use; physical condition; the existence of digital copies elsewhere; copyright status; and the appropriateness of digital reproductions for use and access.

  Because selection criteria for *access* are often different from those for *preservation*, the Library has created a different set of selection criteria for each mission. Each approach focuses on all types of materials in the Library's collections including but not limited to printed monographs and journals, manuscripts and archives, prints and photographs, and audiovisuals.

- NLM Permanence Levels
  Permanence levels assigned to objects in the NLM Digital Repository should be based on the rating system developed by NLM to indicate to users which of its Web documents will be kept permanently available and whether the contents and identifiers of those documents could change over time.
  http://www.nlm.nih.gov/psd/pcm/devpermanence.html

- Security
  The Library will follow the HHS/NIH/OCCS security policies for the NLM Digital Repository.

6.2 Policies to be Reviewed

- NLM DTD book formats
  The NLM Digital Repository will require more expansive, robust DTD encoding for textual materials than is currently available with the NLM Book and Historical Book DTDs. The group recommends the use of the TEI DTD in addition to the NLM Book and NLM Historical Book DTDs for encoding certain text materials. A more detailed justification is included in the report accompanying these requirements.

- NLM Metadata Schemes
  The NLM Metadata scheme should be reviewed to see if it meets current needs. NLM has published an approved metadata schema based on Dublin Core (http://www.nlm.nih.gov/tsd/cataloging/metafilenew.html). The needs of the Repository may require a new schema or additional schemas, more congruent with the existing NLMCommon DTD and the NLMCatalogRecord DTD. (Available at http://www.nlm.nih.gov/bsd/licensee.html)

6.3 Policies to be Developed

- Migration Strategies/Methodologies
  A key area identified that will need policy development is migration strategies and methodologies. Digital preservation is a combination of storing objects in formats that can be migrated, recording appropriate metadata to be able to manage the objects, ongoing monitoring for bit and media degradation and technology obsolescence, and migrating the files to newer hardware and software environments as needed to ensure their continued availability. The preservation planning section of the functional requirements outlines the policy areas that need to be developed including monitoring changes in technology, evaluating the content in the archive, participating in standards development and developing migration plans, including goals, methods and schedules. Policies will also be needed on the decision-making process about whether or when new repository standards will be adopted and when an object can be removed from the repository. See Sections 7.4.4 and 7.5 for more details on specific policy areas.

- Ingest Policies
  Policy development will be needed regarding ingest into the repository, including which staff will be authorized to negotiate submission agreements, minimum requirements for content submitted by producers outside of NLM and maximum embargo periods.

7 Functional Requirements

The functional requirements for the Digital Repository are based on the Reference Model for an Open Archival Information System (OAIS) model and include ingestion of digitized and born-digital materials, metadata generation, data management, archival storage, access, preservation planning, and administration.

An OAIS-compliant repository is an organization of people and systems, which has accepted the responsibility to preserve information and make it available for a Designated Community. OAIS provides a conceptual framework from which the Library can define the core requirements of the NLM Digital Repository.

The OAIS functional model (Figure 2) is separated into six functional entities and related interfaces.

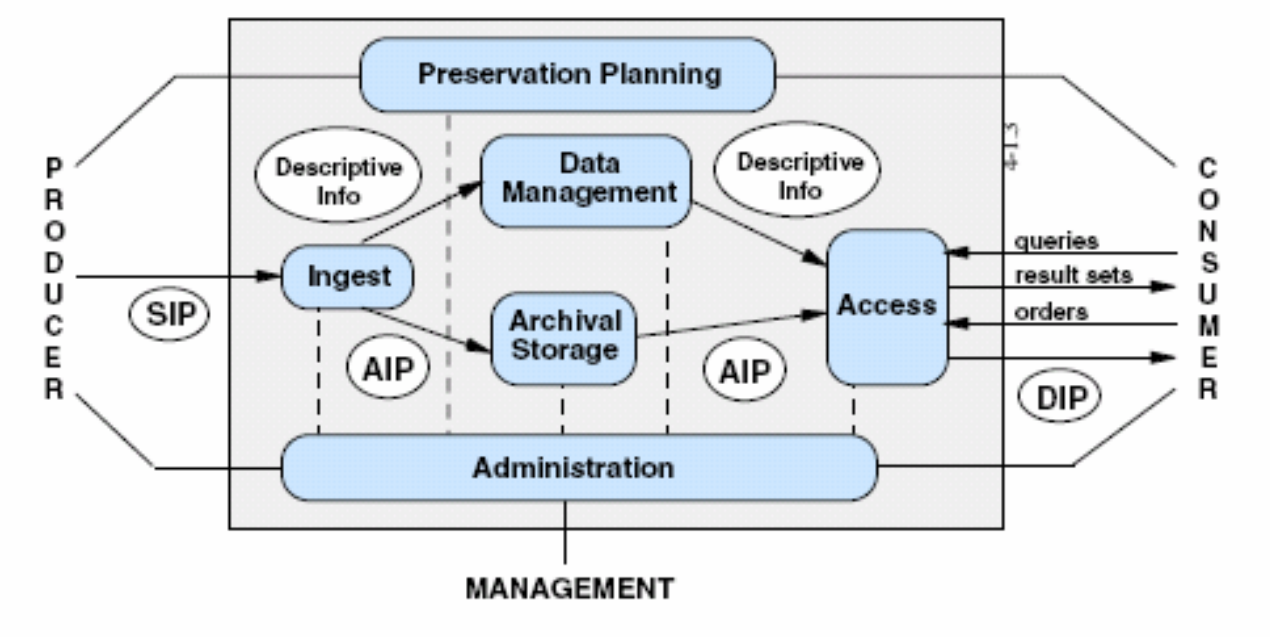

Figure 2 OAIS Functional Entities (OAIS Figure 4.1)

LEGEND:
SIP - Submission Information Package
AIP - Archival Information Package
DIP - Dissemination Information Package

The requirements of the Library are noted for each of the OAIS functional areas. The Library's functional requirements are presented as a series of statements regarding the capabilities needed in the NLM Digital Repository.

For reference purposes the OAIS function descriptions are reproduced using *Italic font*. The requirements of the Library are documented in non-italic font.

A requirement reference number is noted to the left of text that describes the requirement.

7.1 Ingest

Ingest provides the services and functions to accept Submission Information Packages (SIPs) from producers (or from internal elements under Administration control) and prepare the contents for storage and management within the archive. Ingest functions include receiving SIPs, performing quality assurance on SIPs, generating an Archival Information Package (AIP) which complies with the archive's data formatting and document standards, extracting Descriptive Information from the AIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management.

For the NLM Digital Repository, the process of Ingestion is composed of several different functions involved with the processing of selected material to be added to the repository. The actually workflow for this section will vary, depending upon how and when materials are selected and acquired. In general, the Ingest process will include the following phases:

1. Material Selection and Format Preparation
    - Selecting content
    - Working with content producers and donors to establish acceptable formats and transmissions processes
    - Negotiation of submission agreements
    - Review of licensing and registration process

2. Material Submission and Review
    - Authentication of materials
    - Content inventory and virus checking of submitted material.
    - Reviewing submitted materials for acceptance/approval
    - Reviewing materials for correct file formats

3. Ingest of Material
    - Archival Information Package Generation
    - Metadata Generation
    - Coordination of Updates

The repository will have multiple ingest streams, including:

- e-journals and e-books
- harvested Web sites and files
- unpublished manuscript material
- videos
- images
- sound material
- digitized items from NLM's existing collections

The Library may develop other ingest streams in the future. All material that is stored in the repository will enter it via the Ingest function. The function must support the bulk ingest of material as well as ingest of individual objects. The materials submitted must undergo several different quality assurance review processes and potential file conversion in preparation for being stored in the repository.

See Appendix B for the list of formats categorized by level of support.

7.1.1 Receive Submission

The Receive Submission function provides the appropriate storage capability or devices to receive a SIP from the Producer (or from Administration). Digital SIPs may be delivered via electronic transfer (e.g., FTP), loaded from media submitted to the archive, or simply mounted (e.g., CD-ROM) on the archive file system for access. Non-digital SIPs would likely be delivered by conventional shipping procedures. The Receive Submission function may represent a legal transfer of custody for the Content Information in the SIP, and may require that special access controls be placed on the contents. This function provides a confirmation of receipt of a SIP to the Producer, which may include a request to resubmit a SIP in the case of errors resulting from the SIP submission.

In general, the Content Originators with whom NLM negotiates submission agreements (as detailed in Section 7.4.1) are the content producers (those producing published material, i.e. publishers) and content donors. The Library will proactively work with producers and donors to agree on the content, quality and format of digital material and submission schedules. Some Content Providers may have irregular schedules and will submit material on an ad-hoc basis.

The Library will specify preferred formats for the receipt of digital material. However, the NLM Digital Repository must support the acceptance and preservation of material that is provided in non-preferred formats or non-standard formats; these instances should be the exception and are anticipated to be rare occurrences. The Library may want to accept important digital materials in non-standard formats in case we are able to migrate them to a more usable format in the future. For example, scientific data or holographs may be submitted as part of the modern manuscripts program.

Once material has arrived, it must undergo several reviews, including virus checking, format compliance and anticipated content and file type. In some cases, the review and approval step will also include approval of the submitted content. In other cases, the content type may already have been approved and this aspect does not need to be reviewed again.

The repository system must include the ability to record all actions and decisions made concerning the submitted material. The reasons for rejection will be provided back to the Content Provider. In some cases, the provider can then resubmit corrected materials, while in other instances the rejection criteria should prevent the provider from submitting the same or similar material at a later time period.

When submitted material is approved, it will be processed/prepared for ingestion into the repository. This procedure will include the generation of the Archival Information Package (AIP) that ensures that the material conforms to the repository data formatting and documentation

standards. Metadata will be generated, and the components of the AIP will be stored in the system.

Data will be converted to accepted file formats, as needed. Two types of conversions that have been identified are: upgrading the version of file format being submitted (e.g., TIFF5 to TIFF6) or converting from one format to another (e.g., JPEG files to TIFF). A primary goal of the conversion of content for the repository is the preservation of the content. Priority will be given to preserving the intellectual content. The preservation of the original "look and feel" of the original source will be considered and reviewed on a case by case basis.

7.1.1.1	Allows resources to be reviewed before a decision is made whether they should be retained. o Some metadata checks and data conversion may be necessary to allow initial materials to be reviewed o Allows for a definition of the audience for whom the information is intended, who is given access rights, and length of retention Allows for temporary storage outside of the repository (this is optional for certain NLM collections).
7.1.1.2	Supports workflow of digital objects in this order: Receive and track content from producers Normalize content – the conversion into a common format for final ingestion into the repository Validate content based on submitter, expected format, file quality, duplication (e.g., existence of object in repository), and completeness. Review content as a method of the selection process Accept or reject content or file format.
7.1.1.3	A minimal set of identifying information/metadata concerning rejection decisions must be recorded.
7.1.1.4	A minimal set of identifying information/metadata concerning rejected submissions should provide a mechanism to eliminate the need to review some resubmissions. Certain reasons for rejection should prevent a possible resubmission.
7.1.1.5	Allows notification of producers and donors about rejected content.
7.1.1.6	Accepts content from physical media (DVDs, external drives, etc.) or via electronic transfer (FTP) from producers.
7.1.1.7	Accepts content in numerous file types/formats: TIFF images for storage, XML text, audio, streaming video, and PDF.
7.1.1.8	Accepts SIPS in the following formats: XML, MARC, MARC XML, MODS, EAD, DC, TEI, etc.

7.1. 1.9	Accepts the following types of digital objects: Articles, journals, images, monographs, audio files, video files, websites, numeric data, text files, and databases.
7.1.1.10	Converts data to accepted file formats [see Appendix B] as needed. o Examples of conversions may include: upgrading the version of software being submitted; converting one program to another o Conversion needs only to preserve content, not the look, feel and experience of the original
7.1.1.11	Prompts a request for resubmission to the Content Originator if an error of transmission or receipt occurs Error message feature is configurable for all types of errors. System generates an error log. Allows retention of metadata for files rejected but not replaced. Transmission problems allow rejection of entire packages. Other problems will allow rejection of entire package or part of a package, depending on content. Maintains log of messages sent concerning resubmissions.
7.1.1.12	Accommodates a resubmitted SIP in case of errors resulting from the SIP submission.
7.1.1.13	Has the option of sending a configurable automatic system confirmation receipt of SIP to producer based on type of resource, file, producer, etc.
7.1.1.14	Stores and tracks versions of a document/file/image. Links /connections between versions are created and maintained.
7.1.1.15	Assigns a unique identifier to each object. Submitted content may be composed of multiple items, and the system must maintain the relationship between the parent object and any subsequent child objects (i.e., tracking of unique identifiers for child objects and the parent object; possible inheritance of parent object identifier to child objects).
7.1.1.16	Keeps an audit trail of all actions.

7.1.2 Quality Assurance

The Quality Assurance function validates (QA results) the successful transfer of the SIP to the staging area. For digital submissions, these mechanisms might include Cyclic Redundancy Checks (CRCs) or checksums associated with each data file, or the use of system log files to record and identify any file transfer or media read/write errors.

The purpose of Quality Assurance (QA) is to assure the Library that digital material is of sufficient quality to be stored in the NLM Digital Repository. Since manual QA is very time-

consuming, QA activities should be as automated as possible, with the ability to override for manual QA when needed. Manual QA processes must be adjustable depending upon the origin of the file, which could dictate the level of QA required.

7.1.2.1	Performs virus checking on SIP.
7.1.2.2	Validates automatically the successful transfer of the SIP to a staging area using Cyclic Redundancy Check (CRCs) or checksums associated with each data file.
7.1.2.3	Verifies the validity of the submission based on submitter, expected format, file quality, duplication (e.g., existence of object in repository), and completeness.
7.1.2.4	Allows NLM staff to display and perform manual/visual quality control assurance on staged SIPs via a user-friendly GUI interface.
7.1.2.5	Any QA errors shall prompt a request for resubmission and/or internal action. o Internal action may result in creation of a digital object in another preferred format.
7.1.2.6	Allows NLM staff to accept or reject the SIPs at file or batch level.
7.1.2.7	Generates statistical and error reports.
7.1.2.8	Ability to adjust the level of manual quality control needed based on the origin of the file.
7.1.2.9	Keeps an audit trail of all actions.

7.1.3 Generate Archival Information Package (AIP)

The Generate AIP function transforms one or more SIPs into one or more AIPs that conform to the archive's data formatting and documentation standards. This may involve file format conversions, data representation conversions or reorganization of the content information in the SIPs. The Generate AIP function may issue report requests to Data Management to obtain reports of information needed by the Generate AIP function to produce the Descriptive Information that completes the AIP. This function sends SIPs or AIPs for audit to the Audit Submission function in Administration, and receives back an audit report.

Archival Information Packets (AIP) are information packets of content converted to the archive standards identified in Appendix B. Generating the AIPs deals with transforming or repackaging of data to the current NLM archival standards. The system also provides reports containing metadata that may be used in later processes and functions.

The Library will develop policies that govern where the individual components of an AIP are stored. The policy may vary dependent upon a number of factors; for example, the type of digital object, access rights, and conditions of use.

7.1.3.1	Generates AIPs for Repository Storage by transforming SIPs to conform to repository's data formatting standards.

7.1.3.2	Issues report requests to Data Management to obtain reports of information to produce Descriptive Information (metadata) that completes the AIP, as covered in 7.1.1.10 and 7.1.1.13.
7.1.3.3	Capable of interacting with other parts of the system to obtain additional information as needed.
7.1.3.4	Keeps an audit trail of all actions.
7.1.3.5	Converts data as described in 7.1.1.10.
7.1.3.6	AIPs may consist of both masters files and derivatives.

7.1.4 Generate Descriptive Information/Metadata

The Generate Descriptive Information function extracts Descriptive Information from the AIPs and collects Descriptive Information from other sources to provide to Coordinate Updates, and ultimately Data Management. This includes metadata to support searching and retrieving AIPs (e.g., who, what, when, where, why), and could also include special browse products (thumbnails, images) to be used by Finding Aids.

NLM broadly interprets the term Descriptive Information to include descriptive, administrative and structural metadata needed for retrieval and management of digital objects (See Appendix C for metadata definitions.)

Metadata is generated during the Generate Archival Information Package functions.

7.1.4.1	Allows entry and validation of additional metadata (e.g. subject headings, names, dates, “curatorial” descriptive metadata - evaluative information that explains why an object is important, whether it was part of a larger collection (e.g., an exhibit), etc.).
7.1.4.2	Is able to validate specified metadata elements
7.1.4.4	Allows metadata to be stored in a database in a manner that conforms to repository reformatting and linked to their corresponding objects via an identifier. o Basic descriptive metadata will also be stored with the objects (e.g., unique identifier, title and date stored in the TIFF header) so that the objects can still be identified in the event that information in the database is corrupted. o See Appendix D for examples of TIFF header metadata requirements.
7.1.4.5	Is able to recognize required descriptive elements..
7.1.4.6	Recognizes and documents relations among files.
7.1.4.7	Keeps an audit trail of all actions.

7.1.5 Coordinate Updates

The Coordinate Updates function is responsible for transferring the AIPs to Archival Storage and the Descriptive Information to Data Management. Transfer of the AIP includes a storage request and may represent an electronic, physical, or a virtual (i.e., data stays in place) transfer. After the transfer is completed and verified, Archival Storage returns a storage confirmation indicating (or verifying) the storage identification information for the AIP. The Coordinate Updates function also incorporates the storage identification information into the Descriptive Information for the AIP and transfers it to the Data Management entity along with a database update request. In return, Data Management provides a database update response indicating the status of the update. Data Management updates may take place without a corresponding Archival Storage transfer when the SIP contains Descriptive Information for an AIP already in Archival Storage.

The individual components of an AIP may be physically stored in different locations within the repository. The Coordinate Updates function transfers components of the AIP to the relevant storage locations, and maintains a record of the location of the components.

7.1.5.1	Transfers the AIPs to Archival Storage. o Includes a storage request which provides a storage confirmation for the AIP.
7.1.5.2	Transfers the Descriptive Information to Data Management.
7.1.5.3	Sends confirmation notification when transfer to archival storage is completed.
7.1.5.4	Stores the identification information in the Data Management database to support search and retrieval of AIPs in Archive Storage. o Identifiers are the link between Data Management and digital objects.
7.1.5.5	AIPS and Descriptive Information will be put back together when objects are accessed.
7.1.5.6	Keeps an audit trail of all actions.

7.2 Archival Storage

Archival Storage provides the services and functions for the storage, maintenance and retrieval of AIPs. Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfill orders.

7.2.1 Receive Data

The Receive Data function receives a storage request and an AIP from Ingest and moves the AIP to permanent storage within the archive. The transfer request may need to indicate the anticipated frequency of utilization of the data objects comprising the AIP in order to allow the appropriate storage

devices or media to be selected for storing the AIP. This function will select the media type, prepare the devices or volumes, and perform the physical transfer to the Archival Storage volumes. Upon completion of the transfer, this function sends a storage confirmation message to Ingest, including the storage identification of the AIPs.

7.2.1.1	Receives AIPs from Ingest.
7.2.1.2	Moves and adds AIPs to storage within the repository
7.2.1.3	Generates confirmation, statistical and error reports for the AIP process.
7.2.1.4	Keeps an audit trail of all actions.

7.2.2 Manage Storage Hierarchy

The Manage Storage Hierarchy function positions, via commands, the contents of the AIPs on the appropriate media based on storage management policies, operational statistics, or directions from Ingest via the storage request. It will also conform to any special levels of service required for the AIP, or any special security measures that are required, and ensures the appropriate level of protection for the AIP. These include on-line, off-line or near-line storage, required throughput rate, maximum allowed bit error rate, or special handling or backup procedures. It monitors error logs to ensure AIPs are not corrupted during transfers. This function also provides operational statistics to Administration summarizing the inventory of media on-hand, available storage capacity in the various tiers of the storage hierarchy, and usage statistics.

7.2.2.1	Monitors and ensures that AIPs are not corrupted during transfers.
7.2.2.2	Conforms to NLM security policies and requirements to ensure the appropriate level of protection for the AIPs.
7.2.2.3	Keeps an audit trail of all actions.

7.2.3 Replace Media

The Replace Media function provides the capability to reproduce the AIPs over time. Within the Replace Media function the Content Information and Preservation Description Information (PDI) must not be altered. However, the data constituting the Packaging Information may be changed as long as it continues to perform the same function and there is a straightforward implementation that does not cause information loss. The migration strategy must select a storage medium, taking into consideration the expected and actual rates of errors encountered in various media types, their performance, and their costs of ownership. If media-dependent attributes (e.g., tape block sizes, CD-ROM volume information) have been included as part of the Content Information, a way must be found to preserve this information when migrating to higher capacity media with different storage architectures. Anticipating the terminology of section 0, this function may perform Refreshment, Replication, and Repackaging that is straightforward. An example of such Repackaging is migration to new media under a new operating system and file system, where the Content Information and PDI are independent of the file systems. However, complex repackaging and all transformation are performed under Administration supervision by the Archival Information Update function to ensure information preservation.

7.2.3.1	Able to refresh/replace the media on which repository holdings are stored without service interruption, and update corresponding metadata as appropriate.
7.2.3.2	Able to ensure that Content Information and Preservation Description Information (PDI) are not altered.

7.2.4 Error Checking and Disaster Recovery

The Error Checking function provides statistically acceptable assurance that no components of the AIP are corrupted during any internal Archival Storage data transfer. This function requires that all hardware and software within the archive provide notification of potential errors and that these errors are routed to standard error logs that are checked by the Archival Storage staff. The PDI Fixity Information provides some assurance that the Content Information has not been altered as the AIP is moved and accessed. Similar information is needed to protect the PDI itself. A standard mechanism for tracking and verifying the validity of all data objects within the archive may also be used. For example, CRCs could be maintained for every individual data file. A higher level of service, such as Reed-Solomon coding to support combined error detection and correction, could also be provided. The storage facility procedures should provide for random verification of the integrity of data objects using CRCs or some other error checking mechanism.

The Disaster Recovery function provides a mechanism for supplicating the digital contents of the archive collection and storing the duplicate in a physically separate facility. This function is normally accomplished by copying the archive contents to some form of removable storage media (e.g. digital linear tape, compact disc), but may also be performed via hardware transport or network data transfers. The details of disaster recovery policies are specified by Administration.

Error Checking provides statistically accepted assurance to ensure that none of an AIP's components are corrupted via any internal storage function.

The NLM Repository system must be compatible with and conform to the requirements of the NIH Consolidated Collocation Site (NCCS), NLM's off-site back-up facility.

7.2.4.1	Provides statistically acceptable assurance that no components of the AIP are corrupted during any internal Archival Storage data transfer.
7.2.4.2	Performs routine and special data integrity checking such as Cyclic Redundancy Checks (CRC) or checksums for each individual file and generates error reports.
7.2.4.3	Provides disaster recovery capabilities including data backup, off-site data storage, data recovery, etc.

7.2.5 Provide Data

The Provide Data function provides copies of stored AIPs to Access. This function receives an AIP request that identifies the requested AIP(s) and provides them on the requested media type or transfers

them to a staging area. This function also sends a notice of data transfer to Access upon completion of an order.

The Provide Data function provides AIP components to Access.

Material held in the NLM Digital Repository may be subject to restrictions relating to access and conditions of use. See Section 7.6 for issues concerning Access and Rights Management.

7.2.5.1	Provides AIPs to Access to generate Dissemination Information Packages (DIPs).
7.2.5.2	Keeps an audit trail of all actions.

7.3 Data Management

Data management provides the services and functions for populating, maintaining and accessing both Descriptive Information that identifies and documents archive holdings and administrative data used to manage the archive. Data management functions include administering the archive database functions (maintaining schema and view definitions, and referential integrity), performing database updates (loading new descriptive information or archive administrative data), performing queries on the data management data to generate result sets, and producing reports from these result sets.

Data Management provides the services and functions for populating, maintaining and accessing both metadata, which identifies and documents repository holdings, and administrative data used to manage the repository.

7.3.1 Administer Database

The Administer Database function is responsible for maintaining the integrity of the Data Management database, which contains both Descriptive Information and system information. Descriptive Information identifies and describes the archive holdings, and system information is used to support archive operations. The Administer Database function is responsible for creating any schema or table definitions required to support Data Management functions; for providing the capability to create, maintain and access customized user views of the contents of this storage; and for providing internal validation (e.g., referential integrity) of the contents of the database. The Administer Database function is carried out in accordance with policies received from Administration.

7.3.1.1	Maintains the integrity of the Data Management database which contains both metadata and system information.
7.3.1.2	Provides internal validation such as referential integrity of the contents of the database.
7.3.1.3	Creates and maintains schema or table definitions required to support Data Management functions.
7.3.1.4	Has capability to create, maintain, and access customized user views of the contents of this storage.

7.3.2 Perform Queries

The Perform Queries function receives a query request from Access and executes the query to generate a result set that is transmitted to the requester.

All queries will be verified against access rights restrictions.

7.3.2.1	Query requests are received from other functions (for example, Ingest, Access, and Administration).
7.3.2.2	The query request may require data to be sourced from different storage locations.
7.3.2.3	Generates a result set.
7.3.2.4	Allows query requests against all metadata used to manage the repository.
7.3.2.5	Keeps an audit trail of all actions.

7.3.3 Generate Report

The Generate Report function receives a report request from Ingest, Access or Administration and executes any queries or other processes necessary to generate the report that it supplies to the requester. Typical reports might include summaries of archive holdings by category, or usage statistics for accesses to archive holdings. It may also receive a report request from Access and provides descriptive information for a specific AIP.

Management information will be required to support planning, Library management functions, and workflow decisions.

For example, information may be required on the growth in volume of specific types of digital formats across time, on the number of digital objects received on a monthly basis, and patterns of public access.

7.3.3.1	Supports the production of management information reports and statistics.
7.3.3.2	Receives report requests from other NLM Digital Repository functions (for example, Ingest, Access, Administration).
7.3.3.3	Has capability to generate database update confirmation, statistical and error reports to Ingest.
7.3.3.4	Has capability to generate reports in an ad-hoc manner, automatically, or to be triggered by a reporting calendar or by a specific system event.
7.3.3.5	Has capability to generate and provide reports such as summaries of repository holdings by category, usage statistics for access to repository holdings, and descriptive information for a specific AIP.
7.3.3.6	Reports may be specific to a time period or set of time periods.

7.3.3.7	Keeps an audit trail of all actions.

7.3.4 Receive Database Updates

The Receive Database Updates function adds, modifies or deletes information in the Data Management persistent storage. The main sources of updates are Ingest, which provides Descriptive Information for the new AIPs, and Administration, which provides system updates and review updates. Ingest transactions consist of Descriptive Information which identifies new AIPs stored in the archive. System updates include all system-related information (operational statistics, Consumer information, and request status). Review updates are generated by periodic reviewing and updating of information values (e.g., contact names, and addresses). The Receive Database Updates function provides regular reports to Administration summarizing the status of updates to the database, and also sends a database update response to Ingest.

7.3.4.1	Receives updates from other NLM Digital Repository functions (for example, Ingest and Administration).
7.3.4.2	Allows updates to be submitted in batches.
7.3.4.3	Allows online updates to individual records by authorized staff.
7.3.4.4	Has capability to generate and provide management reports and statistics such as summaries of updates by category, user codes, etc.
7.3.4.5	There may be a need to coordinate updates with metadata held in other systems.
7.3.4.6	Keeps an audit trail of all actions.

7.4 Administration

Administration provides the services and functions for the overall operation of the archive system. Administration functions include soliciting and negotiating submission agreements with producers, auditing submissions to ensure that they meet archive standards, and maintaining configuration management of system hardware and software. It also provides system engineering functions to monitor and improve archive operations, and to inventory, report on, and migrate/update the contents of the archive. It is also responsible for establishing and maintaining archive standards and policies, providing customer support, and activating stored requests.

7.4.1	Audits submissions to ensure that they meet archive/repository standards.
7.4.2	Maintains configuration management of the system hardware and software.
7.4.3	Has capability to inventory, report on and migrate the contents of the repository.
7.4.4	Ensures data integrity for version upgrades and format migration.

7.4.1 Negotiate Submission Agreement

The Negotiate Submission Agreement function solicits desirable archival information for the OAIS and negotiates Submission Agreements with Producers. This function also negotiates a data submission schedule with the Producer. It maintains a calendar of expected Data Submission Sessions that will be needed to transfer one or more SIPs to the OAIS and the resource requirements to support their ingestion. This function receives AIP/SIP templates and customization advise from Preservation Planning and sends SIP designs and SIPs to the Audit Submission function as part of the submission approval process. The data submission formats and procedures must be clearly documented in the archive's data submission policies, and the deliverables must be identified by the Producer in the Submission Agreement.

The submission agreement is a formal written agreement between the Content Originator and NLM defining the terms of the content, standards, metadata creation, and transfer schedule of the SIP. The Library will proactively work with Content Originators to agree on the content, quality and format of digital material. Some content originators may have irregular schedules and will submit material on an ad-hoc basis.

A variety of NLM staff will be involved in negotiating submissions. They will be required to work closely with repository systems staff.

SIP design, metadata requirements, and data submission schedules will be negotiated with the Content Originators. These negotiations will be dependent upon the Producer and the nature of the digital material, and agreements may be renegotiated on a periodic or ad-hoc basis.

The Library may schedule ingest of specific types of digital material for specific time periods. For example, ingest of large volumes of web harvested material may be scheduled for specific periods when no other material is being ingested.

Refer to Section 7.1 Ingest, for further information on the actual transfer, validation, and transformation work that occurs in Ingest.

7.4.1.1	Tracks negotiation status, written submission agreements and maintains schedules.
7.4.1.2	Able to add and edit terms of agreement, based on access of level of user.
7.4.1.3	Submission schedules may be set on a periodic or ad-hoc basis.
7.4.1.4	Submission volumes and schedules are managed and monitored.
7.4.1.5	Able to store terms of agreements, and use them to monitor/review/process submissions.
7.4.1.6	Keeps an audit trail of all actions.

7.4.2 Manage System Configuration

The Manage System Configuration function provides system engineering for the archive system to continuously monitor the functionality of the entire archive system and systematically control changes to the configuration. This function maintains integrity and tractability of the configuration during all phases of the system life cycle. It also audits system operations, system performance, and system usage. It sends

report requests for system information to Data Management and receives reports; it receives operational statistics from Archival Storage. It summarizes those reports and periodically provides OAIS performance information and archive holding inventory reports to Preservation Planning. It sends performance information to Establish Standards and Policies. It receives migration packages from Preservation Planning. It receives system evolution policies from the Establish Standards and Procedures function. Based on these inputs it develops and implements plans for system evolution. It sends change requests, procedures and tools to Archive Information Update.

The Manage System Configuration function monitors and maintains the NLM Digital Repository.

7.4.2.1	Monitors functionality of the entire repository.
7.4.2.2	Maintains integrity of system configuration.
7.4.2.3	Audits system operations, performance and usage.
7.4.2.4	Sends requests for system information to Data Management and receives reports.
7.4.2.5	Receives operational statistics from Archival Storage.
7.4.2.6	Summarizes reports and provides repository performance information and repository holdings inventory reports to Preservation Planning.
7.4.2.7	Sends performance information to Establish Standards and Polices.
7.4.2.8	Receives migration packages from Preservation Planning.
7.4.2.9	Develops and implements plans for system evolution.
7.4.2.10	Submits change requests, procedures and tools to Archive Information Update.

7.4.3 Archival Information Update

The Archival Information Update function provides a mechanism for updating the contents of the archive. It receives change requests, procedures and tools from Manage System Configuration. It provides updates by sending a dissemination request to Access, updating the contents of the resulting DIPs and resubmitting them as SIPs to Ingest.

Archival Information Update provides a mechanism for updating Digital Objects (files and metadata) within the repository.

The Library requires the ability to remove AIPs containing objects that have NLM permanence rating of Permanence Not Guaranteed. Objects with Ratings of Permanent may not be removed from the repository, except under specific conditions. An example of these types of circumstances are demonstrated in NLM's existing policy *Errata, Retraction, Duplicate Publication, Comment, Update and Patient Summary Policy for MEDLINE®* (http://www.nlm.nih.gov/pubs/factsheets/errata.html).

This type of action may require the removal of the digital object and its associated metadata, or removal of the digital object and the retention of its associated metadata. Examples of events that could trigger this removal process include:

- Receipt of a Court Order for the removal of material
- Decision to remove outdated digital access copies
- Misrepresentation of ownership by a content originator.

7.4.3.1	Receives change requests, actions and tools from the Manage System Configuration function.
7.4.3.2	Requests DIPs from the Access function.
7.4.3.3	Updates contents of DIPs and resubmits them as SIPs to Ingest.
7.4.3.4	Deletes AIPs from the repository. This may require the removal of the digital object's files, and the retention of associated metadata, or the removal of both the files and metadata.
7.4.3.4	Coordinates the removal of an AIP with the maintenance of metadata held in other systems. Provides an alert that other systems need to be updated.
7.4.3.5	Schedules and performs file migrations.
7.4.3.6	Keeps an audit trail of all actions.

7.4.4 Establish Standards and Policies/Procedures

The Establish Standards and Policies function is responsible for establishing and maintaining the archive system standards and policies. It receives budget information and policies such as the OAIS charter, scope, resource utilization guidelines, and pricing policies from Management. It provides Management with periodic reports. It receives recommendations for archive system enhancement, and proposals for new archive data standards from Preservation Planning. It also receives performance information and archive holding inventories from Manage System Configuration. Based on these inputs, archive standards and policies are established and sent to other Administration functions and the other Functional Entities for implementation. The standards include format standards, documentation standards and the procedures to be followed during the Ingest process. It provides approved standards and migration goals to Preservation Planning. This function will also develop storage management policies (for the Archival Storage hierarchy), including migration policies to assure that archive storage formats do not become obsolete, and database administration policies. It will develop disaster recovery policies. It will also determine security policies for the contents of the archive, including those affecting Physical Access Control and the application of error control techniques throughout the archive.

NLM will address responsibility for operational management, financial management, policy creation, and strategy development for the initial and ongoing development and maintenance of the NLM Digital Repository system. NLM will maintain system evolution paths for all components of the NLM Digital Repository. A plan for the evolution of a component will be developed before the component is implemented. NLM will review existing processes and workflows to determine the extent of change required to manage and preserve digital material.

7.4.4.1	Establishes and maintains NLM Digital Repository best policies, practices and procedures, including those outlined in Sections 6.1 and 6.2.
7.4.4.2	Establishes workflows and business processes for NLM Digital Repository functions.
7.4.4.3	Develops multiple preservation strategies dependent upon, and specific to, the nature of digital material.

7.4.4.4	Receives recommendations for system enhancement and proposals for new procedures from Preservation Planning.
7.4.4.5	Receives performance information and holdings inventory information from Manage System Configuration.
7.4.4.6	Provides approved standards, procedures and migration goals to Preservation Planning.
7.4.4.7	Develops storage management policies including migration policies.
7.4.4.8	Develops database administration policies.
7.4.4.9	Develops disaster recovery policies and procedures.
7.4.4.10	Determines security policies including the application of error control techniques.

7.4.5 Audit Submission

The Audit Submission function will verify that submissions (SIP or AIP) meet the specifications of the Submission Agreement. This function receives AIP/SIP reviews from Preservation Planning and may also involve an outside committee (e.g., science and technical review). The audit process must verify that the quality of the data meets the requirements of the archive and the review committee. It must verify that there is adequate Representation Information and PDI to ensure that the Content Information is understandable and independently usable to the Designated Community. The formality of the review will vary depending on internal archive policies. The Audit process may determine that some portions of the SIP are not appropriate for inclusion in the archive and must be resubmitted or excluded. An audit report is provided to Ingest. After the audit process is completed, any liens are reported to the Producer, who will then resubmit the SIP to Ingest or appeal the decision to Administration. After the audit is completed, a final ingest report is prepared and provided to the Producer and to Negotiate Submission Agreement. Audit methods potentially include sampling, periodic review, and peer review.

The Library will establish a review and audit program for SIPs.

If material is not judged as suitable for collection and preservation then the Producer will be notified. In some cases the Producer may be requested to modify the material, or provide alternate material.

7.4.5.1	Audits data in SIPs or AIPs to ensure that they meet specified requirements.
7.4.5.2	Checks and records required metadata appropriately in the database. Note: Metadata requirements may vary dependent on the information package type (SIP or AIP).
7.4.5.3	Rejects components of information packages that do not meet requirements.
7.4.5.4	Provides audit reports to the Ingest QA function.
7.4.5.5	Keeps an audit trail of all actions.

7.4.6 Activate Requests

The Activate Requests function maintains a record of event-driven requests and periodically compares it to the contents of the archive to determine if all needed data is available. If needed data is available, this

function generates a dissemination request that is sent to Access. This function can also generate orders on a periodic basis where the length of the period is defined by the Consumers or on the occurrence of an event (e.g., a database update).

7.4.6.1	Maintains a record of event-driven events and compares it to the contents of the repository.
7.4.6.2	Generates dissemination requests to Access if needed data are available.
7.4.6.3	Generates orders on a periodic basis (e.g., to update a database).
7.4.6.4	Keeps an audit trail of all actions.

7.5 Preservation Planning

Preservation Planning provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remain accessible to the Designated User Community over the long-term, even if the original computing environment becomes obsolete. Preservation planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge base. Preservation Planning also designs IP templates and provides design assistance and review to specialize these templates into SIPs and AIPs for specific submissions. Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals.

NLM will engage in ongoing preservation planning to ensure the long term availability of digital resources of permanent value.

NLM may implement multiple preservation strategies dependent upon the nature of the digital material within the NLM Digital Repository.

7.5.1 Monitor Designated Community

The Monitor Designated Community function interacts with archive Consumers and Producers to track changes in their service requirements and available product technologies. Such requirements might include data formats, media choices, preferences for software packages, new computing platforms, and mechanisms for communicating with the archive. This function may be accomplished via surveys, via a periodic formal review process, via community workshops where feedback is solicited or by individual interactions. It provides reports, requirements alerts and emerging standards to the Develop Preservation Strategies and Standards function. It sends preservation requirements to Develop Packaging Designs.

The requirements of this function are described in the terms of the OAIS model.
The Monitor Designated Community function:

	The following activities, unless otherwise noted, are performed by NLM or designated committees/teams within NLM.

7.5.1.1	Interacts with repository users and content providers to track changes in their service requirements and available product technologies.
7.5.1.2	Provides reports to the Develop Preservation Strategies and Standards function.
7.5.1.3	Provides requirements alerts to the Develop Preservation Strategies and Standards function.
7.5.1.4	Sends preservation requirements to Develop Packaging Designs.
7.5.1.5	The Repository System keeps an audit trail of all actions.

7.5.2 Monitor Technology

The Monitor Technology function is responsible for tracking emerging digital technologies, information standards and computing platforms (i.e., hardware and software) to identify technologies which could cause obsolescence in the archive's computing environment and prevent access to some of the archives current holdings. This function may contain a prototyping capability for better evaluation of emerging technologies and receive prototype requests from Develop Preservation Strategies and Standards and from Develop Package Designs and Migration Plans. This function sends reports, external data standards, prototype results and technology alerts to Develop Preservation Strategies and Standards. It also sends prototype results to Develop Package Designs and Migration Plans.

7.5.2.1	NLM monitors emerging technologies in order to maintain and improve the architecture.
7.5.2.2	NLM monitors computing platforms (i.e., hardware and software) to identify technologies which could cause system obsolescence.
7.5.2.3	NLM monitors information standards, including metadata standards and data interface standards.
7.5.2.4	The Repository System keeps an audit trail of all actions.

7.5.3 Develop Preservation Strategies and Standards

The Develop Preservation Strategies and Standards function is responsible for developing and recommending strategies and standards to enable the archive to better anticipate future changes in the Designated Community service requirements or technology trends that would require migration of some current archive holdings or new submissions. This function receives reports from the Monitor Designated Communities and Monitor Technology functions, and it receives performance information, inventory reports and summarized consumer comments from Administration. This function sends recommendations on system evolution to Administration. This function also receives external data standards from Monitor Technology and produces profiles of those standards that are sent to Administration as proposals on their potential usage. This function also receives issues from Develop Packaging Designs and Migration Plans in the case of unanticipated submission requirements, and responds with advice to handle the new requirements.

The requirements of this function are described in the terms of the OAIS model.

	The following activities, unless otherwise noted, are performed by NLM or designated committees/teams within NLM.
7.5.3.1	Develops and recommends strategies and best practices to enable the repository to anticipate future changes in the Designated Community service requirements that would require migration of some current repository holdings or new submissions.
7.5.3.2	Develops and recommends strategies and best practices to enable the repository to anticipate technology trends that would require migration of some current repository holdings or new submissions.
7.5.3.3	Receives reports from Monitor Designated Communities.
7.5.3.4	Receives reports from Monitor Technology.
7.5.3.5	Receives performance information from Administration.
7.5.3.6	Receives inventory reports from Administration.
7.5.3.7	Receives summarized consumer comments from Administration.
7.5.3.8	Sends recommendations on system evolution to Administration.
7.5.3.9	Receives external data standards from Monitor Technology.
7.5.3.10	Produces profiles of those standards that are sent to Administration as proposals on their potential usage.
7.5.3.11	Receives issues from Develop Packaging Designs and Migration Plans in the case of unanticipated submission requirements.
7.5.3.12	Responds with advice to handle new requirements.
7.5.3.13	Receives from Administration and reviews disaster recovery policies and procedures.
7.5.3.14	The Repository System keeps an audit trail of all actions.

7.5.4 Develop Packaging Designs and Migration Plans

The Develop Packaging Designs and Migration Plans function develops new IP designs and detailed migration plans and prototypes, to implement Administration policies and directives. This activity also provides advice on the application of these IP designs and Migration plans to specific archive holdings and submissions. This function receives archive approved standards and migration goals from Administration. The standards include format standards, metadata standards and documentation standards. It applies these standards to preservation requirements and provides AIP and SIP template designs to Administration. This function also provides customization advice and AIP/SIP review to Administration on the application of those designs. If this function encounters submissions that are not covered by existing standards and procedures, it can send issues to Develop Preservation Strategies and Standards and receive advice, including new standards, to assist in meeting the new submission requirements.

NLM may request outside review of preservation plans. NLM will test preservation actions in order to determine their effectiveness across a range of digital objects.

The requirements of this function are described in the terms of the OAIS model.

	The following activities, unless otherwise noted, are performed by NLM or designated committees/teams within NLM.
7.5.4.1	Develops new IP designs and detailed migration plans and prototypes to implement Administration policies and directives.
7.5.4.2	Provides advice on the application of these IP designs and migration plans to specific repository holdings and submissions.
7.5.4.3	Receives repository approved standards and migration goals from Administration. Standards include format standards, metadata standards and documentation standards.
7.5.4.4	Applies these standards to preservation requirements.
7.5.4.5	Provides AIP and SIP template designs to Administration.
7.5.4.6	Provides customization advice and AIP/SIP review to Administration on the application of those designs.
7.5.4.7	Requests and receives advice from Develop Preservation Strategies and Standards if submissions are encountered that are not covered by existing standards and procedures.
7.5.4.8	Develops new AIP designs in response to migration goals.
7.5.4.9	Develops prototype software in response to migration goals.
7.5.4.10	Develops test plans in response to migration goals.
7.5.4.11	Develops community review plans in response to migration goals.
7.5.4.12	Develops implementation plans for phasing in new AIPs in response to migration goals.
7.5.4.13	Consults with other functional areas and the Designated Community.
7.5.4.14	Once the migration plan, associated AIP designs, and software have been tested and approved, sends the entire migration package to Administration, which will schedule and perform the actual migration.
7.5.4.15	The Repository System keeps an audit trail of all actions.

7.6 Access

Access provides the services and functions that support Consumers in determining the existence, description, location, and availability of information stored in the OAIS, and allowing Consumers to request and receive information products. Access functions include communicating with Customers to receive requests, applying controls to limit access to specially protected information. Coordinating the execution of requests to successful completion, generating responses (Dissemination Information Packages, result sets, reports) and delivering the responses to Consumers.

A crucial feature of a digital repository created and maintained by NLM will be access to the data by different types of users. These user types, which are detailed in Section 2.0, include:

1. Public Access Level, available to patrons using the Internet both on and offsite.
2. NIH Staff Access level, available to all NIH employees.

3. NLM Staff Metadata Access, available to NLM staff performing cataloging or other indexing of the materials in the repository. This level will allow NLM staff to view materials and add to or edit metadata without changing the object itself.
4. NLM Staff Object Maintenance Access, available to NLM staff who will work with the object itself: adding new files, checking the quality of digital files, manipulating images, performing format conversions and migrations (e.g., from TIFF to PDF, TIFF to JPEG2000, etc.), and investigating problems with the system.
5. NLM System Administration/Programming Access, ultimate rights to the system, required for its management, development, and for assigning appropriate rights to users.

NLM may have to create other levels of access, depending upon the complexity of the copyright and licensing/subscription agreements that are made concerning data placed in the repository system. For instance, licensing agreements may allow certain files obtained by NLM from Content Originators to be available to onsite users only; offsite users may have restricted viewing rights. In addition, some publishers may only allow their material to be served up through NLM's system at a certain interval after its publication. NLM would hold the content on the server but only make publicly available data older than a specified time period.

Access control and rights management will also apply to:

- Audit logs
- System logs
- System files
- Software components

The Library will develop an authentication and access management system to monitor and control access to the NLM Digital Repository and other Library applications. Access within the NLM Digital Repository will be monitored; access rights may be controlled at the individual object level.

7.6.1 Coordinate Access Activities

The Coordinate Access Activities function provides a single user interface to the information holdings of the archive. This interface will normally be via computer network or dial-up link to an on-line service, but might also be implemented in the form of a walk-in facility, printed catalog ordering service, or fax-back type service. Three categories of Consumer requests are distinguished: query requests, which are executed in Data Management and return immediate result sets for presentation to the user; report requests, which may require a number of queries and produce formatted reports for delivery to the Consumer; and orders, which may access either or both Data Management and Archival Storage to prepare a formal Dissemination Information Package (DIP) for on- or off-line delivery. An order may be an Adhoc Order that is executed only once, or an Event Based Order that will be maintained by the Activate Requests function in Administration, and initiated by a dissemination request that may result in periodic deliveries of requested items. The Archival Information Update function in Administration also submits dissemination requests to obtain DIPs needed to perform its update functions. Other special request types are allowed, but are not detailed. This function will determine if resources are available to perform a request, assure that the user is authorized to access and receive the requested items, and notify the Consumer that a request has been accepted or rejected (possibly with an estimate of request cost and an option to cancel the request). It will then transfer the request to Data Management or to the Generate DIP function for execution. This function also provides assistance to OAIS Consumers including providing status of orders and other Consumer support activities in response to an assistance request.

The Coordinate Access Activities function provides an interface layer that will support access functions.

Rights restrictions may limit the level of access that can be provided to Library staff and the public.

The OAIS statement above indicates there is a single user interface to the holdings of the repository. While a single user interface may be desirable, the NLM Digital Repository may provide access to digital objects through multiple interfaces. For example, bibliographic access through LocatorPlus or NLMCatalog with links to the objects, or direct searching of the metadata and/or full-text through a PMC-like interface.
The concept of Access pertains not only to the users of a the system and what level of interaction a user is allowed, but also to the objects stored in the system – what type of actions can be performed on the objects. Material held in the NLM Digital Repository will be subject to the U.S. Copyright Law (Title 17, U.S. Code), as well as potential international copyright agreements. Also, access conditions include software licensing terms and conditions, which may restrict access to specific materials based on user status, location, or time period.

The NLM Digital Repository will be required to hold material that will be embargoed for an extended period (for example, 1 year). Such material cannot be accessed by anyone other than designated staff, and all access or actions relating to the material must be recorded and notified.

Some material may be donated to the Library under strict terms of confidentiality. Access to such material may be subject to a number of restrictions agreed with the donors, such restrictions may relate to access methods and user authorisations. All access or actions relating to the material must be recorded and notified.

The search and retrieval functionality of the repository is an integral part of providing access to the materials stored in repository. This functionality must provide a wide range of abilities in order to serve the user and the various types of materials stored in the repository.

During the development of the functional requirements for the NLM Digital Repository, it has become apparent that NLM will have several repositories – LocatorPlus, PubMed Central (covering journals), and the digital repository currently under investigation. Furthermore, it is possible that the initial NLM Repository may be composed of more than one electronic storage facility, especially in the early development of the repository. This multiple storage facility approach increases the importance of providing a federated searching capability, allowing users to search the multiple storage sites at one time. This functionality should allow users to easily search and retrieve items from any of the NLM storage systems comprising the digital repository.

This section identifies the issues and functionality that different user types should have, the types of access rights and actions that are associated with different object types, and the functionality and features of the search and retrieval system that is required.

User Access:

7.6.1.1	Controls access to data in the repository based on multiple permission levels. These permission levels determine the create/edit/read/delete privileges granted users.
7.6.1.2	General Public Access: Allows restriction to access by the general public, based on licensing terms, time period (embargo periods), location (IP range restrictions), restricted workstation access as well as other possible legal restrictions.
7.6.1.3	NIH Employee Access: Allows restriction to access by NIH employees, based on licensing terms, time period (embargo periods), location (IP range restrictions), restricted workstation access as well as other possible legal restrictions.
7.6.1.4	NLM Staff Metadata Access: Provides an NLM Staff Level allowing access to staff to add or edit descriptive metadata (without changing the files themselves) for production and work related needs such as cataloging, indexing and modifying metadata.
7.6.1.5	NLM Staff Maintenance Access: Provides a restricted NLM Staff Level, allowing specific NLM staff to work with the data itself: adding new files, checking the quality of digital files, manipulating images, editing technical (i.e. preservation) metadata, performing format conversions and migrations (e.g., from tiff to pdf, tiff to jpeg2000, etc.), and investigating problems with the system.
7.6.1.6	NLM System Administration/Programming Access: Provides an NLM Administrative/Programming Level, ultimate rights to the system, required for its management, development, and for assigning appropriate rights to users.
7.6.1.7	Access mechanisms must be sufficiently granular to allow the identification of individual users, in order to maintain audit logs of actions performed by users.

Rights/Data Control of Objects:

7.6.1.8	Access rights and conditions to materials and the directories/folders in which they are kept must provide for one or more of the following basic privileges, either alone or in combination: • Create/Write access • Edit access • Read access • Delete access

7.6.1.9	Access rights may be associated with the metadata relating to an individual object, i.e. the object may be embargoed, with limited internal staff access being available to metadata. Both an object and its related metadata may be embargoed in terms of public access. The system will need to accommodate: • computer interface access • defining access for a specific individual user or group of users • the ability to assign different levels of access to different individuals and groups of users, either in a pre-defined role or in an ad hoc assignment
7.6.1.10	Access rights and conditions of use will be held for each digital object and its related metadata.
7.6.1.11	Access rights and conditions of use will be machine readable and actionable.
7.6.1.12	Access conditions may be specific to a digital object.
7.6.1.13	Access rights and conditions can be inherited from a parent object to any objects designated as a child object.
7.6.1.14	Access rights and conditions can be assigned to an object on an individual basis, or on a group of identified objects at one time.
7.6.1.15	Access status include the following conditions • Free access, where items are freely available via internal or external delivery mechanisms to all users • Restricted access, where access requires permission or satisfaction of some criteria; authorized user access is via an internal or a secure delivery mechanism; conditions can include o User type/status (general public, NIH staff, NLM staff, administrative staff) – access dependent on who the user is in relationship to NLM and NIH o Location – access restricted to specific IP location or physical location (specific workstation) o Time period – access regulated by a designated embargo o Concurrent users – access regulated by the number of concurrent users allowed to access an object at a given time.
7.6.1.16	Objects in the repository are accessible for computational use/machine accessibility such as data mining or automated document retrieval.
7.6.1.17	Retains access to metadata for deleted or retracted content. (Access may restricted to NLM staff only).
7.6.1.18	Allows metadata harvesting by other institutions following the OAI-PMH guidelines.

Search and Retrieval Functionality:

7.6.1.19	Searches interface will be web-accessible and must be Section 508 compliant.
7.6.1.20	Provides metadata searching.
7.6.1.21	Provides full text searching.
7.6.1.22	Provides standard boolean search functions.
7.6.1.23	Provides proximity searching.
7.6.1.24	Provides "more like this" functionality.
7.6.1.25	Search results display should include: • Date sort display • Relevancy ranking display • Author alphabetic display • Source alphabetic display
7.6.1.26	Relevancy ranking should be manipulable via system systems; ideally, user defined settings should also be provided.
7.6.1.27	Image searching should be provided, including: • Figure/description search • Actual image search (like chemical structure search/ draw sample image and find something similar)
7.6.1.28	Provides sound/audio searching
7.6.1.29	Provides federated searching of different repository sites.
7.6.1.30	Advanced search features should include: • Search history • Saved searches • Saved citation lists/bibliographies • Alerts o Functions (topic updates, issue update, cited references, correction notices, etc) o Formats (email, rss, podcasts, others) o Dynamic selection of delivery media without recreating search query • Limits (like LocatorPlus / PubMed functions)
7.6.1.31	System should provide a variety of standard display formats, including the ability for the user to customize the search results display as desired.
7.6.1.32	Alternate search interfaces should be available for searching via alternate mechanisms such as handhelds and PDAs.

7.6.1.33	Provides access to the appropriate copy of the identified item (text, image, video, etc.) for the user.
7.6.1.34	Search result integration with library holdings (like current LinkOut function).
7.6.1.35	Provides quick response time.
7.6.1.36	Allows searching by outside search engines such as FirstGov, Google and Yahoo, according to current NLM protocols and security policies.
7.6.1.37	Allows for external access to other repositories or systems performing web harvesting functions according to the NLM Repository's accepted standards and protocols for automated access (in accordance to NLM robot.txt files and other instructional materials for harvesting and external access).
7.6.1.38	Supports use of multiple languages and non-Roman scripts in search, retrieval and display.
7.6.1.39	Provides access to all versions of digital objects in the repository.
7.6.1.40	Provides system settings and user-defined settings for search functions.

7.6.2 Generate DIP

*The Generate DIP function accepts a dissemination request, retrieves the AIP from Archival Storage, and moves a copy of the data to a staging area for further processing. This function also transmits a report request to Data Management to obtain Descriptive Information needed for the DIP. If special processing is required, the Generate DIP function accesses data objects in staging storage and applies the requested processes. The types of operations, which may be carried out, include statistical functions, sub-sampling in temporal or spatial dimensions, conversions between different data types or output formats, and other specialized processing (e.g., image processing**). This function places the completed DIP response in the staging area and notifies the Coordinate Access Activities function that the DIP is ready for delivery.*

The Generate DIP function accepts a dissemination request, and retrieves the information required to form a Dissemination Information Package.

7.6.2.1	Data returned as a result of a dissemination request may consist of: • one or more digital objects and associated metadata • one or more items of metadata • a response that indicates the User is not authorized to access the material requested. • supports the bulk extraction of files/objects and /or associated metadata.
7.6.2.2	Generation function accepts a dissemination request.
7.6.2.3	Generation function retrieves the AIP from Archival Storage and moves a copy of the data to a staging area for further processing

7.6.2.4	Generation function creates and sends a report request to Data Management to obtain appropriate metadata.
7.6.2.5	The prepared DIP response is placed in the staging area and a message is generated and sent to Coordinate Access Activities that the DIP is ready for delivery.
7.6.2.6	If special processing is required, the Generate function accesses data objects in staging storage and applies the requested processes.
7.6.2.7	Keeps an audit trail of all actions.

7.6.3 Deliver Response

The Deliver Response function handles both on-line and off-line deliveries of responses (DIPs, result sets, reports and assistance) to Consumers. For on-line delivery, it accepts a response from Coordinate Access Activities and prepares it for on-line distribution in real time via communication links. It identifies the intended recipient, determines the transmission procedure requested, places the response in the staging area to be transmitted, and supports the on-line transmission of the response. For off-line delivery it retrieves the response from the Coordinate Access Activities function, prepares packing lists and other shipping records, and then ships the response. When the response has been shipped, a notice of shipped order is returned to the Coordinate Access Activities function and billing information is submitted to Administration.

The deliver response function handles the online delivery of responses to Consumers. The Library will not provide offline delivery services.

7.6.3.1	Display interface is web-accessible.
7.6.3.2	Has export function to provide XML output for batch downloads (similar to Eutilities in PubMed)
7.6.3.3	Allows user to save digital content to a hard-drive, to e-mail document and to save search results.
7.6.3.4	A response request is received from Coordinate Access Activities. The intended recipient is identified, the appropriated transmission procedure determined and the online responses are placed in a staging area in preparation for delivery.
7.6.3.5	Once the response has been sent, an appropriate confirmation message is returned to the Coordinate Access Activities section.
7.6.3.6	Keeps an audit trail of all actions.

8 Metadata Requirements

Extensive redundant storage of metadata whether in the header of the object itself, within the data management system or in multiple access systems, creates problems for keeping updated or changed data in sync. The amount of redundant data stored in the object should be kept to a minimum required for identification of the object.

8.1 Metadata Requirements

See Appendix A for the Minimum Descriptive Metadata Requirements, and Appendix D for an example of the technical metadata in TIFF headers.

8.1.1	System should accept metadata associated with objects in at least the following formats: All NLM DTDs Dublin Core MARC21 MARCXML ONIX MODS EAD TEI
8.1.2	System should have built-in checks on the incoming metadata. Records not containing the minimally defined set of fields should be flagged as problems, either to be returned to the submitter, or sent locally for metadata enhancement.
8.1.3	System should have a user-friendly method of mapping non-standard metadata elements into approved NLM elements.
8.1.4	Once ingested, metadata should be stored in a single common format. This format should be one that ensures against data loss, and allows a variety of access/distribution options, such as all the schemes mentioned above.
8.1.5	System must have the ability to allow for metadata updates.
8.1.6	System must have the ability to search and display metadata, preferably in a user-conformable, human readable display as well as in its native format for machine harvesting and manipulation.
8.1.7	Objects in the repository shall have sufficient technical metadata to assure functionality (e.g. viewing and display) in the present and enable forward migration for future accessibility and use. Technical metadata requirements for objects created by NLM will be developed on a format-specific basis.

Appendix A – Minimum Descriptive Metadata Requirements

Where possible, elements have been assigned names using the terminology from the NLMCatalogRecord DTD (available at: http://www.nlm.nih.gov/databases/dtd/index.html).

Minimal Descriptive Metadata Required From Supplier				
Element	**Required**	**Definition**	**Notes**	**Examples**
TitleMain.Title	M	A name given to the resource		
AuthorList.Author	MA	Name of a person or body associated with the creation of the resource		
PublicationInfo.Publisher	MA	Name of the entity responsible for making the resource available		
DateCreated	M	Date of creation of the original resource, if not originally a digital object		
PublicationInfo.DateIssued	MA	Date the digitized resource was made available		
ResourceInfo	M	Information about the type of resource from a physical aspect, such as resource type, form of issuance, etc.	NLM DTD splits the single DC concept of Format into the two different elements: Resource information and Physical description, with several subelements	Digitized image; Text; PDF file; Streaming video; digitized pamphlet
Physical Description	M	Physical aspects such as extent, size, duration		532 KB; 14 p.; 32 min.
Rights	M	Rights held in and over a resource, such as intellectual property rights or copyright	No current equivalent in NLM DTD	
Language	MA	Language of the resource	Not used for non-textual material (e.g. pictorial images)	
OtherID	M	Identifier supplied by source	Used With Source attribute, giving name of supplier	

Additional Fields Required For Storage In Repository				
Element	**Required**	**Definition**	**Notes**	**Examples**
NlmUniqueID	M	Identifier supplied by NLM		
URL	M	URL where resource is found		
MeshHeadingList.MeshHeading	M*	Subject of the content represented in MeSH		
KeywordList.Keyword	M*	Subject of the content expressed in free text or keywords		
DateRevised	MA	Date of any changes to the digitized object		
PermanenceLevel	M	Extent to which a user can be assured the resource will remain stable and available	NLM currently has 4 permanence ratings	
PermanenceGuarantor	MA	Agency responsible for guaranteeing the availability and stability of the resource		

Key:

M **Mandatory**
MA **Mandatory if applicable**
***** **One of these two fields is required**

Appendix B – NLM Repository Formats

- **Supported:** NLM fully supports the format.
- **Known:** NLM recognizes the format, but we cannot guarantee full support.
- **Unsupported:** NLM does not recognize the format; these will be listed as "application/octet-stream", or Unknown.

MIME type	Description	Extensions	Level
application/marc	MARC	marc, mrc	supported
application/mathematica	Mathematica	ma	known
application/msword	Microsoft Word	doc	known
application/octet-stream	Unknown	(anything not listed)	unsupported
application/pdf	Adobe PDF	pdf	supported
application/postscript	Postscript	ps, eps, ai	supported
application/sgml	SGML	sgm, sgml	known
application/vnd.ms-excel	Microsoft Excel	xls	known
application/vnd.ms-powerpoint	Microsoft Powerpoint	ppt	known
application/vnd.ms-project	Microsoft Project	mpp, mpx, mpd	known
application/vnd.visio	Microsoft Visio	vsd	known
application/wordperfect5.1	WordPerfect	wpd	known
application/x-dvi	TeXdvi	dvi	known
application/x-filemaker	FMP3	fm	known
application/x-latex	LateX	latex	known
application/x-photoshop	Photoshop	psd, pdd	known
application/x-tex	TeX	tex	known
audio/x-aiff	AIFF	aiff, aif, aifc	supported

MIME type	Description	Extensions	Level
audio/basic	audio/basic	au, snd	known
audio/x-mpeg	MPEG Audio	mpa, abs, mpeg	supported
audio/x-pn-realaudio	RealAudio	ra, ram	known
audio/x-wav	WAV	wav	supported
image/gif	GIF	gif	supported
image/jpeg	JPEG	jpeg, jpg	supported
image/png	PNG	png	supported
image/tiff	TIFF	tiff, tif	supported
image/x-ms-bmp	BMP	bmp	known
image/x-photo-cd	Photo CD	pcd	known
text/html	HTML	html, htm	supported
text/plain	Text	txt	supported
text/richtext	Rich Text Format	rtf	supported
text/xml	XML	xml	supported
video/mpeg	MPEG	mpeg, mpg, mpe	supported
video/quicktime	Video Quicktime	mov, qt	known

Appendix C – Glossary of Terms

Administrative Metadata
According to the Metadata Encoding and Transmission Standard (METS), there are four main forms of administrative metadata. 1) Technical Metadata (information regarding files' creation, format and use characteristics); 2)Rights Metadata (copyright and license information); 3) Source Metadata (information regarding the original source from which the digital object derives); and 4) Digital Provenance Metadata (information regarding source/destination relationships among files, including master/derivative relationships between files and information regarding migrations/transformations employed on files between original creation or digitization of an object and its current incarnation).

AIP
See: Archival Information Package

Archival Information Package (AIP)
The version of the information package that is stored in the digital repository. The AIP consists of the digital object and a complete set of metadata sufficient to support the repository's preservation and access services.

Authentication
A process that verifies that an individual, computer, or information object is who or what it purports to be, in order to provide access to material that is restricted by license.

Delivery
The process by which digital content is presented to the user. Delivery of content can be managed by XSLT or CSS style sheets or another interface mechanism for display on the web or via other means.

Descriptive Information
The descriptive, administrative and structural metadata needed for the retrieval and management of digital objects. See: Metadata

Descriptive Metadata
Metadata that serve the purpose of discovery, identification and selection of digital objects which includes elements such as title, author, subjects, etc.

Designated Community
An identified group of potential consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.

Digital Provenance Metadata
Information regarding source/destination relationships among files, including master/derivative relationships between files and information regarding migrations/transformations employed on files between original creation or digitization of an object and its current incarnation.

DIP
See: Dissemination Information Package

Dissemination Information Package (DIP)
The version of the information package that is delivered from the repository to the user in response to an access request.

Donor
An organization or individual providing material to the repository.

DTD
Document Type Definition. A DTD is a formal description of a particular type of document. It sets out what names are to be used for the different types of element, where they may occur, how they can be used, and how they all fit together. DTD were originally used with SGML; and are now used with XML. Elements (categories of information) and tags (the markup that identifies chunks of content as a particular element) are set in the DTD. Usage is also specified, such as vocabulary used to elaborate on element types, whether elements are required, whether elements can contain other elements, and whether they can repeat. A Schema is a newer type of DTD.
The grammar or the construction of a DTD for XML is set by the World Wide Web Consortium (W3C) at http://www.w3.org/TR/REC-xml. Individual DTDs are defined by many organizations in the community.

Dublin Core
The Dublin Core metadata element set is a standard for cross-domain information resource description. It can be used in Web pages to describe their content, or as a minimal metadata standard for cataloging digital files. The standard was developed by the Dublin Core Metadata Initiative. http://dublincore.org/

EAD
Encoded Archival Description. A markup format for describing archival finding aids. The standard was specified by the Library of Congress and the Society of American Archivists. http://lcweb.loc.gov/ead/

Fixity Check
The process of verifying that a file or bitstream has not been changed during a specified period of time.

Harvest
An automated process for identifying and capturing Web materials and/or their metadata for collection or archive purposes.

Header
The header is the first section of a file, which includes metadata embedded by the creator of a digital information resource for description and management purposes. While this is often used to index a file for discovery and retrieval, it is not necessarily displayed as part of the content.

Ingest
The process through which objects are added into the Digital Repository.

IP Address
Unique numerical identifier given to each computer on the network and server on the Internet. The IP address is the address through which you find resources and how data finds its way from a web site back to your computer. A domain of IP addresses is the range of IP addresses assigned to the institution reflected by that domain. A URL is a mnemonic alias for a server's IP address and the location of files in its directory structure.

Metadata
Literally, "data about data," metadata includes data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation. There are literally hundreds of metadata standards specified by national and international organizations. Although often subdivided into categories such as administrative, descriptive and technical metadata, there are no clear lines dividing these categories. See: Descriptive Information.

METS
Metadata Encoding and Transmission Standard. A standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using XML. Can be extended to use elements from other descriptive standards as necessary. The format was specified by the Library of Congress. http://www.loc.gov/standards/mets/

Migration
A set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and ensure their continued availability and usability over time.

MODS
Metadata Object Description Schema. MODS is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. MODS is expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users. http://www.loc.gov/standards/mods//

Normalization
The process of converting digital objects of a particular type (e..g., color images) to a single chosen file format that is thought to embody the best characteristics of functionality, longevity and preservability.

OAI

Open Archives Initiative. Standards for the encoding, harvesting and the construction of repositories of metadata records describing local or remote collections. OAI provides standards and best practice guidelines for the creation of OAI tools by other organizations and institutions. The standard and protocol was developed and is maintained by the Open Archives Initiative Steering Committee. http://www.openarchives.org/

OAI-PMH

OAI-Protocol for Metadata Harvesting (OAI-PMH). Protocols for application-independent interoperability framework based on metadata harvesting, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). Includes standards for data providers who administer systems that wish to expose metadata and service providers that use harvested metadata. http://www.openarchives.org/OAI/openarchivesprotocol.html

OAIS

Open Archival Information System. A reference model and general framework that defines concepts and functions (ingest, management, archival storage, preservation and access) of a digital repository. See ISO Standard 14721
http://public.ccsds.org/publications/archive/650x0b1.pdf

Open Source

A concept through which programming code is made available through a license that supports the users making changes to the code. Any changes are submitted to a group managing the open source product for possible incorporation into the official version. Development and support is handled cooperatively by a group of distributed programmers, usually on a volunteer basis.

OpenURL

OpenURL is a protocol for interoperability between a remote information resource and a local system that offers licensed access. It is a URL that transports metadata or keys to access metadata for the object for which the OpenURL is provided. Electronic journal and database publishers are beginning to support OpenURL access to their resources, and vendors, such as Ex Libris and Sirsi are marketing URL Resolvers (programs that translate requests).
The protocol is specified by the NISO Committee for the Standardization of OpenURL.
http://www.sfxit.com/openurl/ and http://library.caltech.edu/openurl/default.htm

PDI

See: Preservation Descriptive Information

Persistent Identifier

A persistent identifier is name for a resource which will remain the same regardless of where the resource is located. Links to the resource will continue to work even if it is moved. Examples include DOIs (Digital Object Identifiers), OpenURL, Handles, URNs (Uniform Resource Name), and PURLs (Persistent URLs).

Preservation Descriptive Information

Preservation Descriptive information (PDI) consists of data elements needed to ensure the maintenance, migration and long term availability of digital objects in the repository. The OAIS reference model identifies four types of PDI:

- Reference Information: enumerates and describes identifiers assigned to digital objects so they can be referred to unambiguously (e.g., ISBN, URN)
- Provenance Information: documents the history of the digital objects (origins, chain of custody, preservation actions and effects)
- Context Information: documents the relationships of the digital object to its environment (e.g., why it was created and its relationships to other digital objects in the repository)
- Fixity Information: documents authentication mechanisms used to ensure that the content of digital objects has not been altered in an undocumented manner

Producer
An organization, such as a publisher, that produces published material that is deposited within the digital repository.

SIP
See: Submission Information Package

Source Metadata
Descriptive or administrative metadata regarding the original source material (e.g., print version) from which a digital object derives.

SRU
Search/Retrieve via URL. A standard search protocol for Internet search queries, utilizing CQL (Common Query Language), standard query syntax for representing queries. It is a variation of SRW. The protocol is maintained by the Library of Congress -http://www.loc.gov/standards/sru

SRW
A web services implementation of the Z39.50 protocol that specifies a client/server-based protocol for searching and retrieving information from remote databases. It specifies procedures and structures for a client system to search a database provided by a server, retrieve database records identified by a search, scan a term list, and sort a result set. Access control, resource control, extended services, and a "help" facility are also supported. The protocol addresses communication between corresponding information retrieval applications, the client and server (which may reside on different computers); it does not address interaction between the client and the end-user. It is a variation of SRU and maintained by the Library of Congress.

Structural Metadata
Information regarding the internal structure of a digital object and the relationships among its parts (e.g., chapters, sections). Structural metadata enables navigation and presentation.

Submission Information Package (SIP)
Submission Information Package: the version of the information package that is transferred from the content producer to the Library prior to its ingest into the repository.

TCP/IP
Transmission Control Protocol/ Internet Protocol. A standardized suite of network protocols that enables information systems to link to other information systems on the Internet, regardless of their computer platform. TCP and IP are two software communication standards used to allow

multiple computers to talk to each other. The protocol is specified by ISO (International Organization for Standardization) - http://www.iso.ch/ISO/en/ISOOnline.frontpage

TEI

Text Encoding Initiative. A format for representing text documents, designed primarily for encoding their logical structure. The format and rules are expressed as a DTD against which TEI files are checked for conformity to the rules. TEI encoded files include a TEI Header that contains the basic metadata that describes the content (author, title, date, publisher, etc.) in addition to the markup that structures the text itself.
The format is specified by the Text Encoding Initiative Consortium - http://www.tei-c.org/
Software using this format: Note Tab, Word Perfect, and any ASCII text editor can be used to mark up (create) TEI files.

Technical Metadata

Information describing physical attributes or properties of digital objects and how they were created. Some technical metadata properties are format specific (e.g., color space associated with a TIFF image) while others are format independent (e.g., size in bytes).

Unicode

The Unicode Standard is the universal character-encoding standard used for representation of text for computer processing. Unicode provides a unique numeric code (a code point) for every character, no matter what the platform, no matter what the program, no matter what the language. The standard was developed by the Unicode Consortium. http://www.unicode.org/

Z39.50

A protocol that specifies a client/server-based protocol for searching and retrieving information from remote databases. It specifies procedures and structures for a client system to search a database provided by a server, retrieve database records identified by a search, scan a term list, and sort a result set. Access control, resource control, extended services, and a "help" facility are also supported. The protocol addresses communication between corresponding information retrieval applications, the client and server (which may reside on different computers); it does not address interaction between the client and the end-user. The protocol is maintained by the Library of Congress - http://www.loc.gov/z3950/agency/

Appendix D – Technical Metadata

The technical metadata associated with TIFF headers is outlined as follows:

1. **Technical Metadata in TIFF File Headers**
   The NLM repository shall meet the Library's requirement for automatically generated and manually entered technical metadata in TIFF image file headers. NLM will specify which fields are mandatory on a project by project basis.

   a. **Automated technical metadata in the TIFF imager header**
      Technical metadata (along with other pertinent fields) is encoded in the TIFF image header by the scanning software. Most of the mandatory fields in the TIFF image header are internal data, accessed by a TIFF Reader to interpret the image data and render it on the screen.

TIFF Tag	***Field Name***	***Required Field (or Has Default)***	***NISO Technical Metadata***	***Needed for Preservation***
274	Orientation	Y	Y	Y
256	ImageWidth	Y	Y	Y
257	ImageLength	Y	Y	Y
258	BitsPerSample	Y	Y	Y
259	Compression	Y *(D = Uncompressed)*	Y	Y
262	PhotometricInterpretation	Y *(No Default Allowed)*	Y	Y
277	SamplesPerPixel	Y	Y	Y
282	XResolution	Y	Y	Y
283	YResolution	Y	Y	Y
296	ResolutionUnit	Y *(D = Centimeters)*	Y	Y
338	ExtraSamples	Y *(if Field 277 > 3)*	Y	Y

- Two additional fields MIMEType (TIFF) and ByteOrder (Big- or Little-Endian), required by *NISO Z39.87-2002*, are retrievable from the first four bytes of the image file.
- SegmentType (Strip vs. Tile) of the TIFF image is inferred from other TIFF header fields.
- ExtraSamples provides Opacity or Transparency information for images with Alpha channels.