Digital Repository Policies 0.1

Introduction

The mission of the Computer History Museum is to preserve and present for posterity the artifacts and stories of the information age. To accomplish this mission, the Museum must be able to preserve, curate and interpret digital collections that form a significant and rapidly growing component of the Museum's holdings.

Digital preservation can be defined as managed activities and processes necessary to ensure both the long-term maintenance of the byte stream and continued accessibility of its contents.

Digital Repository policies will help translate aspirations and goals into operational plans for the Computer History Museum (CHM). Operational plans help staff members work effectively, through documented policies and procedures for routine tasks. With clear policies and procedures for most situations, staff members spend less time seeking answers for routine matters. This frees up resources for new initiatives and complex problem solving.

CHM is committed to an OAIS (Open Archival Informational System) compliant digital repository infrastructure, following current digital repository best practices. Documented policies and procedures can help the Digital Repository achieve "trusted" status, according to assessment frameworks like PLATTER and TRAC[1].

Digital Repository policies are intended to be consistent with policies for the CHM Collections and Exhibitions department to create a holistic approach to and view of all CHM collections, regardless of format. I have numbered this early version of Digital Repository policies Version 0.1. While it is possible to create some policies based on best practices while the Repository is in the planning stages, these policies will need to be reviewed and fine-tuned based on the experience of implementing the CHM Digital Repository. Therefore, these policies are offered in the spirit of guidelines for the prototype. As the prototype Repository is tested, the policies should be adjusted, and the areas that are dependent on implementation decisions completed.

[1] PLATTER is the Planning Tool for Trusted Electronic Repositories from digital preservation Europe, http://www.digitalpreservationeurope.eu/platter/. TRAC is Trustworthy Repositories Audit & Certification produced jointly by the U.S. National Archives, Council on Research Libraries & OCLC, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

Policy areas

Figure 1shows the components of the Digital Repository system within the Computer History Museum infrastructure. Digital repository policies are required for each function to the right of the “SIP” process, as well as for the Digital Repository data store within the storage and archival area. Policies that govern digitization and digital production are needed, but are out of scope for this document, which focuses on core Digital Repository functions. Policies for generating and making access copies of digital content available are also out of scope.

Overall administrative policies will clarify roles and responsibilities, insure sustainability, and provide contingency plans should the CHM undergo a change in organizational status or decide to discontinue support for the Digital Repository.

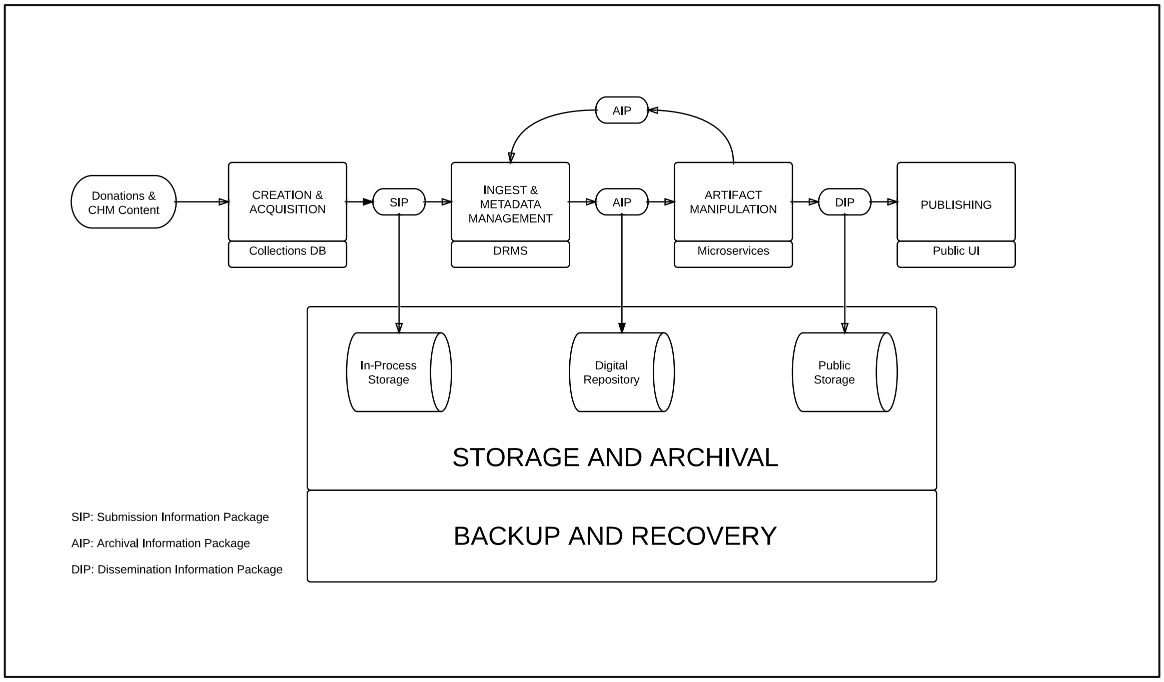

Figure 1: Digital Repository System

This document includes policies in the following areas:

- Administration
- Ingest (Submission)
- Archiving
- Archival Storage, Backup and Recovery
- Interfaces and Access

Please see Appendix B: CHM Digital Repository Glossary at the end of this document for topics referred to throughout the document and other digital repository documentation.

Administration

Authority

The Digital Repository (DR) program is the responsibility of the Museum's Collections and Exhibitions department. The Repository depends on the support, technical infrastructure, and knowledge provided by the Museum's Information Technology department. The DR will be administered by a core team currently composed of the Digital Media Archivist, IT Manager, and Software Curator, with the Director of Collections acting as the team leader. The team will meet at least quarterly to review activities and suggest changes and alternations. Changes identified by the core team, that affect the Museum on a policy level (as determined by the Director of Collections with input from the team) will be presented to the VP of Collections and the CEO.

Purpose

The Digital Repository will provide digital preservation services for digital objects produced by CHM and selected born digital objects that have been acquired by the Museum as part of its permanent collection. Of special concern is preserving high definition video from CHM's media group, currently part of the Collections and Exhibitions Department, as well as preservation worthy digital surrogates of the permanent collection within the context of CHM's collection policies. Please see Appendix A: Records Lifecycle, for a review of when a digital object is ready for archival preservation services.

Donations

When accepting digital objects into CHMs permanent collection, intellectual property rights and legal issues must be considered. Except in extraordinary circumstances, CHM must be granted ownership and/or license rights that allow for preservation, duplication and access. Ownership or rights should be granted through a deed of gift or similar legal contract with the donor or creator. Donated digital objects must meet criteria for selection that is consistent with the CHM mission and collection policies.

There are many issues to weigh in all donation offers. If the donation includes software, does the donor have the right to assign copyright or license to the CHM? If not, is the material vulnerable enough that the CHM would consider preserving it until the rights can be determined or negotiated? If the donation includes personal archival material, are there restrictions on dissemination of the material? Complex or uncertain rights, licensing, or access issues can add significantly to processing time for digital collections. Therefore it is important to balance the intrinsic value of digital donations that are encumbered by rights or access conditions with the level of resources that managing them will require.

In extraordinary situations, donated material may be of significant intrinsic value and may be accepted, even when rights are uncertain (i.e. orphan works), or when access to the material must be restricted. In these cases, restrictions must be documented so access policies can be set correctly when the material is ingested into the repository. CHM policies should comply with limitations to copying and storing, redistribution and use, or any other actions regulated by intellectual property law. Rights policies and procedures for the Digital Repository should be consistent with rights policies and procedures for digital exhibit artifacts, as documented in the Guidelines for Documenting Digital Exhibit Artifacts (27-Apr-2010), Section 3.

Although collecting and making malicious software (malware) available for study may fit with the CHM mission, storing software that has the potential to be harmful is not aligned with the Digital Repository goal to function as a trusted digital repository. Files will be checked for viruses on ingest, and will be rejected if infected. Alternative approaches to preserving malware for study would be to keep the code on quarantined machines, or to rely on printed versions of the code for study.

Deaccessions from the DR will follow the policy laid out in the *Collections Management Policy*. Collections wide acquisition procedures are currently being revised to document the digital provenance and operating environment of digital objects upon being accessioned into the collection. These procedures will help document the authenticity of the digital object over time.

Financial Support and Sustainability

Enduring preservation of digital resources requires substantial and ongoing financial commitments over time – potentially more so than for traditional [paper based] materials. Digital preservation is dynamic; responses to technological obsolescence or media decay must be taken more quickly and the life expectancy of a preservation treatment is shorter because the technologies used are evolutionary. Consequently preservation strategies must be periodically monitored and reassessed as the technological environment that supports standards, protocols, and formats, etc. evolves. [2] The fragility of digital artifacts requires their constant attention and monitoring. Without this level of support the financial cost are even greater when "rescuing" digital artifacts that that have been left to degrade.

[2] Yale University Library Digital Preservation Policy, page 5. Retrieved February 27, 2012 from http://library.yale.edu/iac/DPC/final1.html.

CHM in making a commitment to preserve our digital artifacts to the same standards as our physical artifacts thereby assume the responsibility of making commensurate resources available. It is estimated that initial development is $2.5 million over a 3 to 5 year period including ingest of all current and acquired digital objects and creation of a fully integrated and multi-featured public user interface. Normal digital preservation activities will include ongoing costs for upgrades and monitoring of the technical infrastructure, estimated at $75,000 annually amortized to account for obsolescence of technical systems every 5 to 7 years. Staffing will require a minimum of 1 FTE. We spend over $50,000 for annual maintenance on our off-site storage facility, providing appropriate care to our permanent physical object collection. The DR will also require resources on an appropriate level to guarantee the preservation of digital objects.

Because technical systems are inherently ephemeral and community standards are revised over time, CHM must make a commitment to provide learning opportunities for members of the digital repository team. Likewise the members of the digital repository team have a responsibility to share learning and information within the cultural heritage community.

In the event that the CHM digital preservation mission is redefined, or the CHM ceases to exist in its current form, the CHM will make every effort to insure that the digital artifacts preserved in the CHM DR are transferred to another cultural heritage organization in accordance with our current collections policy for all permanent collection artifacts.

Ingest (Submission)

Responsibility for Ingest Process

The Digital Media Archivist manages the Digital Repository ingest workflow.

Determining Priority for Ingest

CHM-produced content, including high definition video files, and preservation worthy digital surrogates of CHM artifacts should have first priority for DR ingest. The ingest pipeline should be robust enough to accommodate current donated digital content as well as a steady stream of legacy digital collections. The Digital Media Archivist manages the collection ingest queue in consultation with the Director of Collections. Priorities for the DR are:

1. CHM-produced video (oral histories, lectures, exhibit video, etc.);
2. Vulnerable legacy material on physical media such as magnetic tape, floppies, etc. that have undergone curatorial review and been accepted into the permanent collection;

3. Vulnerable new donations on physical media (same formats, same review as above);
4. Stable new donations on any medium accepted into the permanent collection;
5. Stable legacy material on any medium accepted into the permanent collection.

The DR prototype will provide an opportunity to measure ingest throughput requirements, identify processing bottlenecks, and tune systems and processes.

File Formats and Validation

While the ingest workflow should include format validation using a tool such as DROID, JHOVE or FITS[3], permanent collection content that has been deemed preservation-worthy should not be rejected, even if the file format cannot be identified, or if a file of a known type fails validation. Instead, the report from the validation tool should be preserved with the object in the Digital Repository database or data stream as technical metadata so that appropriate levels of preservation service can be applied to the object.

Known file types and validated files may be eligible for migration services whereas the CHM policy will be to preserve unknown file types at the bit-preservation level.

After prototyping, file types for migration and preservation service levels will be added to this policy document. Below is a preliminary list of preservation file types produced by CHM that are migrations candidates[4]:

- Moving image: MOV (Quicktime), AVI, DV (videocassettes)
- Still Images: TIF, JPG
- Text: PDF/A, DOC, TXT, RTF
- Sound recording: AIFF, MP3, WMA, WAVE, RMID

The ubiquity of the following file types assumes their longevity because they are either open source, well understood, or have community support for their migration into the foreseeable future:

- Moving image: MOV (Quicktime), AVI, DV (videocassettes)
- Sound recording: MP3, WAVE
- Still Images: TIF

[3] DROID -Digital Repository Object Identification is a software tool developed by the Nation Archives to perform batch ID of file formats. JHOVE – JSTOR/Harvard Object Validation Environment, IDs file formats as well. FITS – File Information Tool Set, created by Harvard, identifies, validates and extracts file format metadata and goes a step beyond JHOVE.

[4] The Library of Congress provides a compendium of file format descriptions including sustainability factors.

- Text: PDF/A, TXT, RTF

Virus Checking

CHM policy is to check files for viruses as soon as they are brought into the CHM environment. Virus checks should be run again each time a file is moved from one file system to another. Information about the virus-checking event (time and result) should be captured as part of the preservation metadata.

Metadata Treatment

Metadata treatment policies for digital collections are aligned with Collections department metadata policies. Just as vulnerable analog collections may receive minimal processing to enable storage in and retrieval from a warehouse style climate-controlled facility, digital collections may be ingested into the Digital Repository for preservation. Once stabilized, they can be retrieved for more comprehensive treatment later, if a use case for access arises, or if resources for full cataloging become available. Minimal descriptive metadata for ingest should include a unique identifier, lot id, title, creator (if applicable), creation date(s), rights statement, and restrictions statement (if applicable).

Note: There is no need to wait to ingest legacy collections such as bitsavers.org into the repository until each object receives full cataloging. Instead, digital collections can be handled on a collection, "series", or file level for ingest, in a similar fashion to the way unprocessed archival collections are handled in the collections database.

Archiving

Object Processing and Maintenance

The Digital Repository preserves well understood file formats for CHM produced content, and file formats that may be unknown, such as legacy software or games. As a result, levels of preservation service vary by file type. DR format migration services can be applied to currently supported file formats produced in house, while bit preservation service is the only service available for unfamiliar file formats.

When files are preserved at the bit level, the Digital Repository insures that the files can be retrieved as they were deposited, but does not provide forward migration services. Emulation environments are also out of scope for the Digital Repository.

Metadata Standards and Format

The Digital Repository uses Dublin Core as the descriptive metadata format as consistent with current Collections policy. Levels of descriptive metadata treatment are aligned with metadata treatment policies for all collections. Vulnerable digital collections may be processed for preservation with minimal descriptive metadata.

Technical and preservation metadata may be stored in the Dublin Core metadata format, or in METS (metadata encoding and transmission standard). After prototyping this document will be updated with the technical and preservation metadata formats depending on what the Digital Repository software supports. [expand based on what software solution supports]

File Format Migration

The Digital Repository most likely will provide file format migration services for CHM produced digital content. The Collections and Exhibits department will regularly review file formats that have been ingested to determine if any are at risk for obsolescence and to plan forward migration to insure that objects can be used when they are retrieved from the repository.

Records will be collected and maintained of all administrative actions taken on a digital object so that all versions created from the original object ingest will be clearly delineated. This will create clear provenance to authenticate the versions of all objects in the repository.

Data Integrity Protection and Audits

A museum guiding principle is the ability to authenticate the artifacts in its custodianship. Confidence in the authenticity of digital objects over time is particularly crucial owing to the ease with which alternations can be made. Data integrity checks, audits and donor information concerning provenance guarantee for future generations the authenticity of the digital objects in CHMs stewardship.

Regular checksum generation provides baseline data integrity checking in the Digital Repository. Checksums should be generated for objects when they are ingested, whenever they are moved from one file system to another, whenever there is a media migration, and whenever an object is retrieved from the preservation system.

Checksum policy is a topic that generates much discussion whenever a group of digital preservation professionals gather. CHM will need to balance best practices against resource constraints when setting data integrity protection and audit policy. Beyond the key events listed in the above paragraph, the CHM should implement a level of data integrity checking it can afford by looking for automated checksum tools that can be run against the Digital Repository on a regular or ongoing basis. CHM will create a procedure for handling checksum error reports. [expand based on storage hardware solutions]

Versioning

On occasion, a digital object that has been ingested will need to be replaced, to correct errors that have been made in the digitization process, to add files that were inadvertently omitted, or to replace files that were included by mistake. In these cases, the CHM should follow emerging best practices for versioning that are supported by the Digital Repository software. This will include retaining the older and original version of the object and ingesting a replacement.

Archival Storage, Backup and Recovery

Archiving Policy

The Digital Repository may consist of a combination of spinning disk and archival tape storage, to be determined by CHM staff in consultation with the Digital Repository Storage Consultant. Although it is desirable to keep first copies of preservation data on disk for ease of integrity checking, if required from a cost perspective, it is acceptable to use tape storage. Using disk for files that may be accessed more frequently, and tape for infrequently retrieved objects is also an acceptable solution. In addition to regular backup of the files on disk, the Digital Repository will include at least two archival copies of the content, each stored in a separate geographic location. The infrastructure undergoing testing (as of Aug. 2012), includes tape & disk backups at our off-site storage facility. At our main facility in Mountain View there will be a WORM archive, working disk space plus a tape & disk back-up.

Hardware and Media Migration

Hardware and media migration are more a function of technology watch than of policy per se. Policy does dictate that the Digital Repository Team should monitor Digital Repository hardware and media life cycles, anticipated to be about 5 years. At least a year before a platform or medium comes to the end of the support period, the team should plan a data migration. These planned migrations also provide opportunity for system-wide integrity checks.

Disaster Recovery

CHM's disaster recovery policies might involve negotiating a pro bono agreement with a hardware vendor that would be willing to get a basic system up and running until the full Digital Repository will be re-built within a week from archival data that has been stored off-site.

Because the CHM Digital Repository provides long term preservation services for cultural heritage materials rather than critical access to health or financial records, the CHM Digital Repository disaster recovery plan does not require immediate 24/7 access. When disaster strikes, CHM's limited resources are unable to provide a mirrored site and redundant hardware for an immediate rebuild.

Security

The Computer History Museum is committed to insuring physical safety and security for all the digital content it preserves. This security is provided by limiting access to the physical facilities that house servers and other equipment where digital content is stored. In addition, access to the DR is restricted to authorized users, and protected by information technology industry standard security protection such as firewalls. [expand based on storage hardware solutions]

Interfaces and Access

User Interfaces

For the prototype repository, the direct user interface will be limited and restricted to internal authorized users. User accounts will be maintained by the Collections department. Interfaces with internal and external systems will be enabled through Application Programming Interfaces (APIs), controlled by CHM Information Technology staff. At a minimum, the Digital Repository will import data from and may export data to the collections database, and any metadata stores that enable access through the CHM website.

Search Functionality/Discovery

Digital Repository search functionality will be determined by the level of metadata provided for the objects and by the Digital Repository software that is selected. Digital Repository search functionality may be different from the CHM website search functionality. Metadata for publicly accessible objects may be exported from the Digital Repository to an aggregated metadata store that is exposed to Google for crawling and is used by the CHM website for public searching. These workflows are yet to be determined and are out of scope for the prototype of the Digital Repository project. [It is anticipated that by 2015 advanced search functionality and a publicly accessible online user interface to the Museum's artifact collection will be implemented.] [expand based on software solution functionality]

Delivery and Emulation

The prototype repository will not deliver digital objects directly to an end user through the CHM website. Before objects are ingested into the repository, any derivatives or surrogates that would normally be needed for access will be produced and stored in file systems that are separate from the Digital Repository. Only if an access copy needs to be replaced will an object be retrieved from the Digital Repository to create a replacement copy or derivative.

Although the repository may preserve software designed to make objects usable, such as players, viewers, or emulation environments, enabling use at this level is out of scope for the Digital Repository prototype project.

Appendix A: Records Lifecycle

All records whether in analog or digital form follow a lifecycle, beginning at creation, moving to active and semi-active use, and finally disposition - either archival preservation or destruction/deletion. For the Digital Repository ingest should consist of records that are ready for disposition, meaning they are not edited, changed or accessed at any regular interval, they are not current records.

Current Records are those that continue to be actively used through ongoing changes, revisions, editing or repurposed with sufficient frequency to justify keeping them in the office of creation.

Archival Records are no longer used in the day-to-day course of business, but which may be transferred out of the creating office for preservation and occasionally used for legal, historical, or operational purposes.

Within the CHM context, all donations to the permanent collection have reached the archival records stage and both the creators or current owners and CHM curators have determined that these artifacts should be permanently retained without alternation due to their historical value.

CHM produced content follows the same records lifecycle. Because of the need for frequent access for the day to day needs of current records, it is unrealistic, in the repository prototype, to expect the digital repository to provide 24/7 access for media productions. Once the records have reached archival records status they will be ingested into the digital repository. Before ingest digital repository managers will provide assistance and expertise to the media team throughout the digital object lifecycle, including procedures to make ingest as timely, efficient and easy as possible.

Appendix B: CHM Digital Repository Glossary

access

Providing object representations to users; terms and conditions of granting permission to use data resources and collections in an archive.[1]

Archival Information Package (AIP)

An information package composed of a digital object bitstream, a UID, and metadata comprised of technical, preservation, administrative, and structural.

audit

Tracking all the interactions with records within an electronic system so that any access to the system can be documented as it occurs for the purpose of preventing unauthorized actions in relation to the records.[1][2]

authenticity

Insuring that an object is what it purports to be; the practice of verifying that a digital object has not changed or is not corrupted.

bulk operations

Operations on a large set of objects as opposed to granular operations on a single (or small set) of objects.

dark archive

An archive that is inaccessible to the public.[3]

digital object

The encapsulation in digital form of an intellectual or aesthetic work including content and description. [2]
Simple digital object: a file
Complex digital object: multiple files which, together with their structural metadata, make up a single object.

digital preservation

The managed activities necessary for ensuring the long-term retention and usability of digital objects.[3]

digital surrogate

Representation of an object available to end user during discovery process; often a lower resolution version of the object.

Dissemination Information Package (DIP)

An information package comprised of the bitstream and corresponding descriptive metadata, derived from an AIP that is delivered to the user in response to a request.

discovery

The process of finding and identifying objects of interest by an end-user.

emulation

The imitation of a computer system, performed by a combination of hardware and software, that allows programs to run between incompatible systems. Or, the ability of a program or device to imitate another program or device.[3]

encapsulation

Process of grouping information into an information object (data and metadata) along with instructions on how its contents should be accessed and interpreted to enable future interpretation; may be used to combine data (content), associated metadata, and a viewer to render the combination into a single object.

fixity checking

Ensuring that an object is unchanged from its accepted state using strategies including redundancy, error-correcting codes, message digests, and hashing (using a mathematical algorithm against data to produce a numeric value that is representative of that data).

format analysis

Identification of the logical format of an object.

identity

Distinguishing one an object from another, typically by methods of persistent naming and actionable resolution strategies.

identity service

Service that supports minting new identifiers, binding associated identifier to its referent and resolving to retrieve a referent for a given identifier.
UID: unique identifier
OID: object identifier
DOI: digital object identifier

immutability

Reliably preventing changes to data that could occur through erasure, corruption, or loss.

ingest

The process by which a digital object or metadata package is taken in by a different system than the one that produced it.

integrity

The validity of data, either its intellectual validity (content has not been changed) or its physical validity (bits have not been altered e.g. during migration from one format to another, or during transmission from one system to another, or by a virus)[4] interoperability The ability of multiple systems, using different hardware and software platforms, data structures, and interfaces, to communicate, exchange, and share data.

metadata Structured information that facilitates discovery, management and preservation of an object, a collection of objects, or a constituent part of an object such as an individual content file. Digital objects that do not have sufficient metadata or become irrevocably separated from their metadata are at greater risk of being lost or destroyed. Ephemeral, highly transient digital objects will often not require more than descriptive metadata. However, digital objects that are intended to endure for long periods of time require metadata that will support long-term preservation.[3]

migration The transfer of digital objects from one hardware or 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 to retain the ability for clients to retrieve, display, and use them in the face of constantly changing technology. Migration includes refreshing as a means of digital preservation, however, it is not always possible to make an exact digital copy of a database or other information object and still maintain the compatibility of the object with a new generation of technology.[3]

normalize To migrate a digital object to a chosen logical format; to conform metadata to a structured vocabulary or other formalization.

provenance Documentation of the history (including origin, source, and changes) and chain of custody of a digital asset since it originated.

replication service Methods for making authentic copies of objects or files for the purposes of preservation.

repository Storage domain designed and operated to provide access to, as well as retain, protect, and preserve authentic digital objects.

retrieval Process of locating and obtaining objects from storage (digital or physical).

scalable
A process or system that initially supports a small number of operations or objects but can be expanded without complication to support much larger numbers.

security
Enforcing appropriate use of services and content using methods such as identity and role management, and physical and virtual separation of systems.

Submission Information Package (SIP)
An information package composed of digital object bitstream, lot ID & metadata in the form of administrative, checksums, virus/malware test pass.

transformation
Transcoding of digital object representations from existing available forms to standard formats (from proprietary to open source, for example). Also called format migration.

trustworthiness; trusted
In the context of electronic records, trustworthiness often implies that the system is dependable and produces consistent results based on well-established procedures.[5]

validation
A process to check one or more aspects of a submission for schema errors, file format problems, and ingest parameter inconsistencies that might affect its suitability for preservation. Results of a validation may include any combination of structural analysis information, warning messages, or fatal errors that prevent an object from being ingested.[3]

version
Successive change in an object; a variant based on changes made to the content (data) or the metadata over time, as the content evolves and changes over its active state.

version control
Techniques, especially in an automated environment, to control access to and modification of documents and to track versions of a document when it is revised.[5]

[1] RaivoRuusalepp (2003). AHDS Digital Preservation Glossary. Retrieved March 13, 2012 from http://www.ahds.ac.uk/preservation/preservation-glossary.pdf

[2]

[3] California Digital Library. Glossary. Retrieved March 13, 2012 from http://www.cdlib.org/inside/diglib/glossary/?field=glossary&action=search&query=preservation

[4] José Luis Borbinha & Fernando Cardoso (2000). NEDLIB Glossary. Retrieved on March 15, 2012 from http://www.kb.nl/hrd/dd/dd_links_en_publicaties/nedlib/glossary.pdf

[5] Society of American Archivists. A Glossary of Archival and Records Terminology. Retrieved from March 15, 2012 http://www.archivists.org/glossary/term_details.asp?DefinitionKey=2579