erpa

Information Society Technologies

Ingest Strategy

September 2004

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK

Introduction

Increasing quantities of information are presented as digital objects. Storing these objects is no simple matter as it is being found that storing digital objects to ensure access and authenticity is far more complex than for their paper counterparts. Correspondingly, there is a growing need to understand the scope, perspectives and factors of ingest of digital objects into storage areas. This guidance document is intended to introduce ingest and its role in the development of a digital repository system. The appendix contains a companion guide and checklist when defining and/or selecting an ingest strategy, presenting a survey of the factors required for consideration. This document does not provide a definition of current strategies as these are still maturing.

Additionally, the "Digital Preservation Policy", "Cost Orientation", and "Selecting Technologies" Guidance documents may be of use to the reader.[1]

Context

The role and function of ingest, also known as acquisition, is specified in the CCSDS Reference Model for an OAIS[2]. Selection and appraisal are functions that are separate to ingest and have not been defined here. The OAIS Model serves as a complete discussion of the functions a repository system should contain. In the context of this document, ingest refers to the services and functions required to:

- Accept a submission package;
- Prepare the contents for storage and management;
- Perform qualitative assurance on the submission packages; and
- Create and derive descriptive and technical metadata required for the coordination and management of information within a system.

Role of Ingest

Storing a digital object upon a controlled medium and ensuring the bitstream can be persistently discovered and accessed can extend the lifespan of an object. To enable this in practice, ingest is required to identify and record the appropriate semantic and syntactic properties of the object. An ingest strategy dictates the procedures and mechanisms required to acquire and retain information at a specified quality. A clear ingest strategy will clarify the objectives and goals of an organisation, ensuring a continuing understanding of intent that will persist over time.

An ingest strategy should never be considered complete. As new formats, technologies and standards are defined, it is logical that the ingest strategy will grow in both complexity and sophistication. Producing a strategy that can incorporate such changes while retaining the extendibility and manageability of the original design should be a key goal. Throughout all developments the activities must remain financially sustainable in the long term. To cater for this, regular reviews of the strategy in relation to technological change should occur.

Prerequisites

An ingest strategy should always be developed in collaboration with several components of a repository system. However, this can cause ambiguities as to where the boundaries of ingest lie. The methods of appraisal and selection can influence

[1] See http://www.erpanet.org/guidance/.
[2] Reference Model for an Open Archival Information System (OAIS); CCSDS; January 2002; Page 4-1 http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf.

the ingest strategy; affecting who the responsibility of depositing information lies with and the quantity of metadata required. The architecture of the repository must be developed in conjunction with the ingest strategy as the process of ingest includes the creation of metadata dependant upon the implemented architecture.

The appraisal strategy will determine part of the information deposited alongside a digital object at submission. The appraisal strategy requires clearly defined boundaries of curation and selection, alongside a definition of the audience for whom the information is intended, and who is given access rights. This will provide information specifying the length of retention and method of selection for deposit. It is inadvisable to ingest objects unnecessarily and retain them for longer than is required as both ingest and retention are costly and time consuming processes[3], and in some cases may be illegal.

Registration assists resolution of the Intellectual Property Rights (IPR) issues that can result from storing, copying and distributing digital objects. The intention of registration is to effectively form a contract with the originator of the object, so documents must be created accordingly. It is important to clearly identify the persons or organisations that can be held accountable for actions. This information must also be stored in perpetuity alongside the digital object. Note that this procedure is an activity distinct from ingest.[4]

Prior to developing an ingest strategy the legal requirements must be clarified. Must digital objects be retained for legal purposes? Can the authenticity and integrity of the object be guaranteed before, during and after ingest?[5]

The roles and responsibilities of the staff must be determined. If loss of data occurs, it should be clearly stated who holds responsibility and whether this is an individual or an organisation. If an individual, that person must be aware of the position they retain.[6]

Selecting a preservation strategy should not be a prerequisite for determining the ingest strategy. An ingested object should not be restricted or tied to a particular long-term preservation strategy. Rather, the stored metadata should be both generic and comprehensive enough to remain useful for newly developed alternate preservation strategies.[7]

---

[3] Resources of note here include the InterPARES Project “Appraisal Task Force Report”, http://www.interpares.org/book/index.cfm; and The University of Michigan Libraries “Selection Criteria for Traditional and Electronic Resources”, http://www.msu.edu/~wellsat/draft_report.doc.

[4] Noteworthy resources include: Intellectual Property Rights Lessons from the CEDARS Project for Digital Preservation, http://www.leeds.ac.uk/cedars/colman/CIW03.pdf; and the Cedars Guide to IPR, http://www.leeds.ac.uk/cedars/guideto/ipr/.

[5] Authenticity resources include: Authenticity in a Digital Environment, http://www.clir.org/pubs/reports/pub92/pub92.pdf; Again, InterPARES (http://www.interpares.org) has produced documents of note.

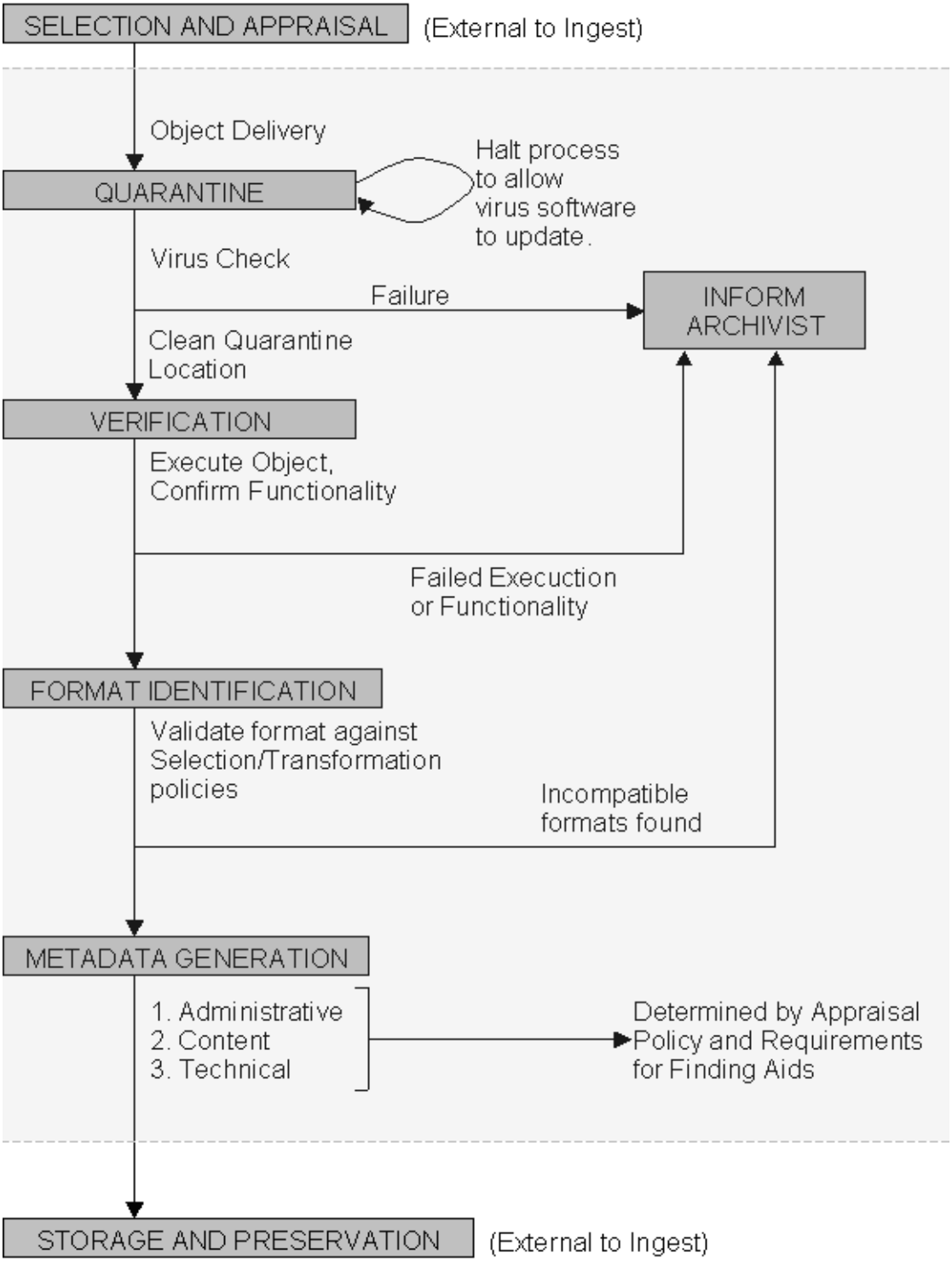
[6] Roles are briefly discussed in Lavoie's “Incentives to Preserve Digital Materials”, http://www.oclc.org/research/projects/digipres/incentives-dp.pdf; Maggie Jones' “Roles and Responsibilities of Collecting Institutions in the Digital Age”, http://www.nla.gov.au/nla/staffpaper/npomj.html discusses relevant topics.

[7] There is a high quantity of available literature on digital preservation strategies. A nice introduction can be found at http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf.

Functions of Ingest

The following diagram details a generic workflow of the functions required by an ingest process. A key goal of the ingest strategy will be to automate these procedures where possible. The requirements of automation will be reduced by the introduction of format and representation information registries.

Figure 1: The functions of ingest in relation to the adjacent layers



As stated in *Context*, **selection and appraisal** require a distinct policy. For most parts of this workflow, it will be desirable to automate as much as possible. At each stage of the procedure, controls should be developed to inform the 'archivist'[8] in the event of a problem or failure. After delivery, **quarantine**[9] allows an object to be stored temporarily while it is virus-checked. The system will also ensure that its virus detection software is updated before it is run. **Verification** ensures that the object is compositionally complete and executes correctly, and **format identification** that the object is composed of well formed files in formats previously identified. The generated **metadata** must be produced at a high level of quality. The quantity of metadata will be determined by the **appraisal** policy (recording what the object contains and where it originated), and the requirements of finding aids (enabling future searching and identification of objects). Finally, the **storage and preservation** activities require appropriate and separate policies.

Factors

An ingest strategy is required to store information:

- efficiently and accurately;
- at a minimum cost.
- with minimal human effort;
- for the maximum required duration;
- for the correct and specified audience.

The benefits and effectiveness of the strategy, if well designed, should be immediately apparent and also allow suitable risk assessment to be performed on the contents of the system.

Costs

Determining accurate cost models is very difficult at this early stage of the development of data archiving. It is already clear that well-defined strategies and best practice guidelines will reduce the cost of the ingest procedure. A recent consultancy review states that the most significant cost faced by data managers is trying to clean up digital resources that should have been cleaned at the point of creation.[10] There will be many fixed costs in a complete preservation strategy, such as system and infrastructure development, and these are difficult to avoid. The variable and ongoing costs will be generated through the creation of new data and the maintenance of existing data, and it will be here that well-defined strategies will assist in cost minimisation.

Conclusion

Developing a successful ingest strategy requires implementing the functions necessary for ingest and also ensuring dependencies with non-trivial strategies and systems. In this rapidly changing environment, attention should be paid to any new and evolving processes. Remaining aware of developments and implementing

[8] We use the term to denote someone who may not necessarily be an archivist, but rather one who is managing and looking after the archive or repository.
[9] See the National Archives of Australia, http://www.naa.gov.au/recordkeeping/preservation/digital/digital_repository.html.
[10] See Hendley, Tony; Comparison of Methods & Costs of Digital Preservation, 1998. http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/hendley-report.pdf.

suitable changes as and when necessary, but not rashly, will assist ensuring the information required is acquired at an appropriate quality and cost.

Ingest Strategies - Appendix

The appendix to this document presents a survey of factors that should be addressed when developing an ingest strategy.

Standards

Standards range from the structure of persistent identifiers[11], technical and bibliographic metadata schemas[12], through to fully featured archival systems[13]. Standards overlap and it can be difficult to select that which is most suitable. There are few widely accepted standards relating to ingest and metadata creation[14]. It is important to select a standard that is closest to the goals of the organisation while retaining flexibility for modification.

STANDARDS	Bibliographic and Technical Metadata	Which standard is right for the goals of the organisation?
		How will the fields of the metadata standard be qualified?
		Will the implemented metadata remain platform independent?
		Can the implemented model be extended without rendering existing entries obsolete?
		Can interoperability of the metadata be achieved?
	Coverage of Schema	Is the implemented metadata model receptive enough to cater for variations in content (e.g. TIFF)?
		Does the schema enable suitable properties of data content files to be recorded (e.g. font, resolution)?
		Does the schema enable suitable properties of binary files to be recorded (e.g. OS and Hardware configuration)?
		Can suitable semantic metadata be recorded to verify the object has been correctly represented without reference to the original?
		Is the metadata recorded in a machine-readable form?
	Identification	How will objects be identified uniquely and persistently? In this developing field, what temporary solution can/will be implemented?
		How will metadata be linked to the digital objects?
		Do the metadata allow appropriate database query and identification?

As further experience in metadata usage is gained standards will develop and mutate, and concurrently implementation and best practice guides will become

[11] See the Digital Object Identifier, http://www.doi.org; and Persistent URL, http://www.purl.org. Also see the ERPANET seminar held on this topic: http://www.erpanet.org/events/2004/cork/index.php.
[12] These include but are not limited to: the PREMIS group http://www.oclc.org/research/projects/pmwg/; NLNZ, http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf; and Dublin Core, http://dublincore.org/
[13] Such as the OAIS, http://www.rlg.org/longterm/oais.html
[14] The exception would be the OAIS model, ibid.

available. These may reduce the need for an institution to qualify fields and provide methods for interoperability. Regardless, this will remain a challenging area of the system design.

Content data formats, such as raster image formats, contain properties that should be recorded to formally identify the object's content. Conversely, binary executables contain very few properties that can be explicitly recorded. It may be more useful to place an emphasis on the operating system and hardware configuration under which this executable should run.

Development of persistent identifiers is ongoing.[15] It is likely to be some time until we see a standardised system. It is infeasible to wait for some standard to be developed and accepted. It will be important to ensure that any system used can be altered, re-referenced or replaced. To reduce risk, it may be possible to implement multiple identifier systems in parallel while waiting for a community decision on which system is most suitable.

Systems, Methodologies and Technologies

Systems, methodologies and technologies can also rapidly change. Code modularity and strong system documentation is important. This is especially true in long-term situations where staff change, helping avoid the reverse engineering of obsolete, unfamiliar technology. Appropriate development paradigms should be used, manuals detailing the systems, and handbooks outlining the usage processes should be created and updated as necessary.

SYSTEMS, METHODOLOGIES, TECHNOLOGIES	Object Transfer and Delivery	Are the transfer and delivery routes secure and standardised for high risk-value objects?
		Should obsolete formats be accepted for acquisition?
	Quarantine	Has a secure temporary quarantine location been created?
		Does this location get cleaned regularly?
		Is the virus checking software up to date?
	Archival and Storage	Where will the logical objects reside in relation to the database?
		Is the storage medium replaceable and scalable over time?
		Have backup and contingency plans been developed?

It is possible to transfer objects electronically (e.g. SCP, FTP, or download) or conventionally (physical transfer of portable media). Submission procedures will vary accordingly. Increased automation and a transfer of responsibility to the depositor may be achieved by developing an electronic system, though displaced responsibility can introduce errors in metadata creation. Conventional methods may result in issues arising from obsolete media formats, though the proliferation of information stored on portable media means that this area should not be neglected.

[15] PADI offers an aggregation of information sources, http://www.nla.gov.au/padi/topics/36.html. Interestingly, the ERPANET Report on its Persistent Identifiers Seminar highlights the gulf between research and actual implementation. http://www.erpanet.org/events/2004/cork/index.php.

Quarantine[16] allows an object to be stored temporarily ensuring virus detection software to be updated before it is run. Suitable security and locations must be provided for this system, and appropriate virus checking software must be obtained. Introducing viruses into a large database management system can have devastating effects and this step should not be overlooked.

Laws and Policies

Can the authenticity and integrity of the object be guaranteed during and after ingest? Have methods for enabling this been designed and implemented in a way that is flexible within a changing environment?

LAWS AND POLICIES	Verification	Can the authenticity of the object be verified?
		Can digital signatures be used to verify the long-term authenticity of the object?
		Can the completeness of the object be verified? Can this process be automated?
		How will hash functions verify the integrity of the content?
		What level of granularity must hash functions operate on?
	Provenance	Has an appropriate provenance record been created?
		How will alternate and outdated versions be stored and identified?
		How does the strategy distinguish between a correction to a metadata record and an alteration to a provenance record?
		Can information be updated in a metadata record? For what period of time after creation?
	Quality Control	How can the quality of the ingest process be audited?
	Documentation	Do the implemented policies have appropriately detailed documentation?
		Will this documentation remain meaningful to new users over long periods of time?
	Accountability	Who is to be held accountable for management and control of the system?

In many situations digital objects must be retained for legal reasons. The authenticity of the object must be verifiable. Techniques such as digital signatures must be investigated for this role. Appropriate security measures must ensure the objects cannot be compromised by unauthorised sources. Hash functions can verify that the

[16] See the National Archives of Australia, http://www.naa.gov.au/recordkeeping/preservation/digital/digital_repository.html.

byte stream has not been altered. Policies must determine what granularity of the object the hash function must operate on.[17]

An appropriate provenance trail must be recorded. This must dictate whether to retain outdated versions of an object, how these can be identified and how they can be used. Can errors in a metadata record be corrected? Who can authorise this? What of the original information should be retained? Is the strategy flexible and secure enough to allow and enable our evolving understanding of content?

The lack of suitable documentation on file formats and existing systems is a significant obstacle of digital preservation. This should be prevented from reoccurring in any newly implemented system. Documentation of policies is equally important. By understanding why information is being retained, more emphasis can be placed on achieving this goal.

Appropriate quality control and accountability systems must be installed. How is the correct running of the system verified regularly and over long periods of time? Suitable cost effective and efficient backup systems must be installed to prevent loss of data.

Practices

Best practice can only be discovered through experience. As institutions develop ingest strategies, guidance based on institutional experience will be published. It is feasible to set up a manual ingest strategy. However, this is likely to result in a low quality system from the introduction of human errors, increased by the time consuming, technical and repetitive nature of the work. Automation of the system is required to reduce such errors and increase quality.

PRACTICES	Automation	Can the electronic transfer and registration process be automated?
		Can the authenticity of the object be automatically verified?
		Can the completeness of the object be automatically verified?
		Can the identification of file formats be automated?
		Can the identification of technical and bibliographic metadata be automated?
		Can the metadata entries be validated automatically?
		Can the correct operation of the system be verified automatically?
		Can the quality control procedure be automated?

Automation is integral to lowering the costs of an ingest strategy. However, automating a system is still technically complex and can require high initial and recurring costs. These arise from modifications that must occur as new technology and standards are integrated into the system. While these costs may appear to be high, it is likely that a successfully automated system will have a lower resultant cost as the number of technicians required is drastically reduced. It may not however, be possible to fully automate the system. If this is the case, the strategy must detail in

[17] See Lynch's “Canonicalization” paper, http://www.dlib.org/dlib/september99/09lynch.html; and “Authenticity and Integrity in a Digital Environment”, http://www.clir.org/pubs/reports/pub92/lynch.html; also see the InterPARES Authenticity Task Force, http://www.interpares.org.

which area it is most cost beneficial to automate and the feasibility of this investigated.

Integration with external systems may further reduce the institutional workload requirements. For example, work is underway on developing a global digital format registry.[18] If this was to meet the initially drafted goals, the identification of formats and extraction of technical properties may be more realistic for many institutions.

People

When building a delicate system, staff can cause both success and failure. Employing suitable staff requires a clear definition of the roles required and the skill set needed. Successful management and reviews of staff can ensure a system does not deteriorate over time. In a field with such a rapid development cycle, training should not be neglected.

PEOPLE	Tasks	Different people may be required for different tasks. Example tasks, for areas surrounding ingest, include: Selection, Identification and Ingest Technician, Maintenance and System Design, Quality Control and Management.
	Skills	The skill set of an ingest technician will depend on the level of automation. With high automation, the skill set can be minimal and require simple data entry. Manual ingest requires a high understanding of many file formats to correctly extract the technical properties. As manual ingest is a time consuming role for a relatively highly skilled person, emphasis should be placed on automating and streamlining the process.
	Training	Rapid developments in both IT and digital preservation strategies require continuous training for all those involved.
	Experience	Can vary. The level of automation and the user interface of the system may reduce the experience required by technicians.

Ingest technicians must be suitably trained according to the system. A highly automated system is preferable as it reduces the number of technicians required, the time taken per ingest, and the workload and technical knowledge base of the technician.

Regular training is essential in any fast moving field. Appropriate funds and time allocation are required to inform staff about developments in policies and strategies, system technologies, and advances in delivery formats. This is particularly important for manual ingest technicians, who must understand the properties of file formats under investigation.

[18] See http://hul.harvard.edu/gdfr/.

Bibliography:

ERPANET; Resource Gateway; http://www.erpanet.org

Preserving Access to Digital Information; Resource Gateway; http://www.nla.gov.au/padi/

The OCLC/RLG Working Group on Preservation Metadata: A Metadata Framework to Support the Preservation of Digital Objects; June 2002; http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

National Library of New Zealand: Metadata Standards Framework- Preservation Metadata (Revised); June 2003; http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf

OCLC/RLG PREMIS Working Group: PREMIS (PREservation Metadata: Implementation Strategies); 2003; http://www.oclc.org/research/projects/pmwg/default.htm

Global Digital Format Registry; http://hul.harvard.edu/gdfr/

Reference Model for an Open Archival Information System (OAIS); CCSDS; January 2002; http://ssdoo.gsfc.nasa.gov/nost/isoas/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

The Long Term Preservation of Authentic Electronic Records; InterPARES Project; http://www.interpares.org/book/index.cfm

Selection Criteria for Traditional and Electronic Resources; The University of Michigan Libraries; http://www.msu.edu/~wellsat/draft_report.doc

Intellectual Property Rights Lessons from the CEDARS Project for Digital Preservation; http://www.leeds.ac.uk/cedars/colman/CIW03.pdf

Cedars Guide to Intellectual Property Rights; http://www.leeds.ac.uk/cedars/guideto/ipr/

Authenticity in a Digital Environment; CLIR; 2000; http://www.clir.org/pubs/reports/pub92/pub92.pdf

Incentives to Preserve Digital Materials, Brian Lavoie; 2003; http://www.oclc.org/research/projects/digipres/incentives-dp.pdf

Roles and Responsibilities of Collecting Institutions in the Digital Age; Maggie Jones; 1995 http://www.nla.gov.au/nla/staffpaper/npomj.html

Migration: Context and Current Status; Digitale Bewaring Testbed; http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf

How digital records are transferred to the long-term digital repository; Andrew Wilson; 2003 http://www.naa.gov.au/recordkeeping/preservation/digital/digital_repository.html

Canonicalization: A Fundamental Tool to Facilitate Preservation, Clifford Lynch, 1999, http://www.dlib.org/dlib/september99/09lynch.html

Authenticity and Integrity in the Digital Environment; Clifford Lynch; 2000; http://www.clir.org/pubs/reports/pub92/lynch.html