William M. Lazorchak. The Ghost in the Machine: Traditional Archival Practice in the Design of Digital Repositories for Long-Term Preservation. A Master's Paper for the M.S. in L.S degree. May, 2004. 139 pages. Advisor: Helen Tibbo

This research paper explores the ways in which traditional paper-based archival principles and practices, based on the requirements of a knowledge system built on physical forms (papyrus, parchment, paper), are being applied to the development of repositories designed explicitly for the long-term preservation of digital materials.

Though debate remains active, the archival community has gradually coalesced around a set of high-level principles and practices generally agreed as representative of the core values of archival activity: the sanctity of evidence; the preservation imperative; the primacy of the record; respect des fonds, original order and provenance; and hierarchy in records and their collective description. These traditional archival principles and practices are defined, then translated into digital repository architecture designs through an analysis of the Open Archival Information Systems reference model (OAIS). Areas of active research on this subject are examined in a set of case studies.

Headings:

Preservation of Library Materials

Electronic Data Archives - Conservation and Restoration

Information Systems - Special Subjects - Research

Open Source Software - Evaluations

Archives - Aims and Objectives

THE GHOST IN THE MACHINE: TRADITIONAL ARCHIVAL PRACTICE IN THE DESIGN OF DIGITAL REPOSITORIES FOR LONG-TERM PRESERVATION

by
William M. Lazorchak

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

May 2004

Approved by

Helen Tibbo

Contents

Acknowledgements

Special thanks to my advisors and employers: Dr. Helen Tibbo, Dr. Gary Marchionini and Dr. Paul Jones.

Thanks to the amazing staff at ibiblio for technical assistance, and to all the interviewees for their graciousness in sharing their knowledge, and their valuable time: Mackenzie Smith; Dale Flecker; Patricia Galloway; Carl Fleischhauer; Bruce Ambacher; and Paul Conway.

Very special thanks to my wife Madelyn and son Graham for going away and coming home.

1: Introduction

This research paper explores the ways in which traditional paper-based archival principles and practices, based on the requirements of a knowledge system built on physical forms (papyrus, parchment, paper), are being applied to the development of repositories designed explicitly for the long-term preservation of digital materials.

Though debate remains active, the archival community has gradually coalesced around a set of high-level principles and practices generally agreed as representative of the core values of archival activity: the sanctity of evidence; the preservation imperative; the primacy of the record; respect des fonds, original order and provenance; and hierarchy in records and their collective description. These traditional archival principles and practices are defined, then translated into digital repository architecture designs through an analysis of the Open Archival Information Systems reference model (OAIS). Areas of active research on this subject are examined in a set of case studies.

> Today, information technologies that are increasingly powerful and easy to use, especially like those that support the World Wide Web, have unleashed the production and distribution digital information. Such information is penetrating and transforming nearly every aspect of our culture. If we are effectively to preserve for future generations the portion of this rapidly expanding corpus of information in digital form that represents our cultural record, we need to understand the costs of doing so and we need to commit ourselves technically, legally, economically and organizationally to the full dimensions of the task. Failure to look for trusted means and methods of digital preservation will certainly exact a stiff, long-term cultural penalty.[1]

Archivists are cultural heritage professionals in the role of caretakers of arguably the most important knowledge of our age. The custodial responsibility for society's memory is imbued with immense responsibility, and fraught with technical and social difficulties, yet archivists have regularly been up for the challenge:

> The archival community ...work[s] to define and promote the social utility of records and to identify, preserve, and provide access to documentary heritage regardless of format. Archival holdings are noncurrent organizational records of enduring value that are preserved by the archives of the creating organization ...The archival perspective brings an evidence-based approach to the management of recorded knowledge. It is fundamentally concerned with the organizational and personal processes and contexts through which records and knowledge are created as well as the ways in which records individually and collectively reflect those processes.[2]

Traditional archival principles and practices were based on the requirements of a knowledge system built on physical forms: papyrus, parchment, paper. Even as these forms became intellectually complex (photographs, audio recordings), they retained a physicality that grounded archival practice fully in the realm of the senses. Digital materials are now largely invisible, while their complexity increases exponentially. Seen as a flickering ones and zeros mediated by a computer screen, these digital materials have lost the sense of historicity that previously radiated from the most profound works of collective memory. Digital materials, still imbued with evidentiary and informational value, are abstracted from the archivist's physical sensations, causing a profound, elemental rethinking of the archivist's creed.

While mankind has kept records from the beginning of society, modern archival principles and practices came to fruition in the second half of the 20th century, stimulated (certainly in the United States) by the explosion of government records generated by World War II. The first task of archivists in the post-war period was to attain physical and intellectual control over the rapidly proliferating records in their possession. Digging out from under the mountain of paper proved to be difficult, but the intellectual disinterring imposed a self-analysis upon the profession, leading to the devise of common practices for archival science, spearheaded by T.R. Schellenberg of the National Archives and Records Administration (NARA).[3] The contemporary digital explosion, and the transforming qualities of digital materials, has similarly exacted a profound sense of self-reflection by the profession. Dr. Patricia Galloway of the University of Texas has argued that archival terminology based on paper-based materials cannot be translated into the world of digital materials,[4] but the core principles and practices of archival science remain important to the study of digital information, and the application of these principles and practices to digital systems promises to have significant value for society. The modern archival challenge is the testing and application of appropriate traditional principles and practices to a consideration of digital materials.

Digital preservation has been defined as the managed activities necessary for ensuring both the long-term maintenance of a bytestream of digital data and the continued accessibility of its contents.[5] Digital preservation makes significant demands on archivists, many of whom are technically and fiscally ill-equipped to deal with the complexities of the issue, but the archival community's experience in managing physical materials implies an expertise that can be brought to bear on materials in digital form. Thus, archivists have joined the research community of academics and university alliances, standards-making bodies such as the World Wide Web Consortium (W3C),[6] and open source and forward-thinking commercial software developers to attempt to gain some control over the management and preservation of digital materials:

> Archivists have accepted the challenge, and the responsibility, to collect, preserve, and protect this fragile, constantly changing record of who we are and what we do. From the smallest shoe box stored in a hall closet to the voluminous National Archives, documents and memorabilia require special handling and an awareness of the materials particular value and possible use. Archivists are trained and experienced to deal with the various questions and problems which arise in the preservation of such material. They are able to bring a specialized perspective and an informed interpretation to decisions concerning the material's worth and

> usefulness. They are familiar with the nature and characteristics of all types of human documentation—from ancient Egyptian papyrus to contemporary computer e-mail. And archivists understand that within each document is a drama, behind each letter or photograph is a person. They are safe-guarding not the residue of our culture, but the immediate and permanent resources which will define who we are and explain what we did, for posterity.[7]

The core archival principle of *responsible custody* is manifested in the digital arena in the form of the digital repository, but despite several decades of awareness of the acute digital preservation needs, the development of trusted repositories in the digital world which echo the physical spaces of traditional archives has been marked by small, tentative steps. An analysis of the architecture of modern digital preservation systems must begin with the identification of centrally established doctrines from the brick-and-mortar world of archival practice.

2: Methodology

The list of traditional archival principles and practices examined in this paper has been developed from an exhaustive review of the literature on traditional archival practice. After introducing and defining the essential archival principles and practices, I will explore the efforts that have been made in the information science community to articulate those principles and practices in digital repository system design. These high-level mappings are best represented by the Open Archival Information System (OAIS) Reference Model (RM), an international standard which has become an important model for the information science community. I provide an overview to this model, noting the political processes surrounding its development that introduced archival values into its design. The discussion of the OAIS RM provides the context in that the previously defined archival principles and practices become mapped to functional requirements for digital repository architectures. In addition to the OAIS RM, I will discuss other paths of research that have introduced archival influences on system design, and include mappings of these components to repository architectures.

I will then use the case study method to look more explicitly at a set of pre-operative or operative digital repository architectures strongly influenced by the OAIS RM in their repository design. These case studies are informed by personal interviews conducted with significant contributors to these systems. The individual cases were chosen by analyzing the available literature on OAIS-compliant repositories, and then selecting the small subset of institutions that were implementing, or attempting to implement, these repositories. A selection of the key individuals working on these projects was then contacted. The names and contact information were gathered from publicly available documents supporting these research efforts. Initial contact with the repository staff was made through electronic mail, and follow-up appointments were then scheduled, either in person, or by telephone. The interviews covered a broad range of subject areas related to the design of digital repositories, but were concentrated on analyzing the ways in the which the OAIS reference model had been applied to that particular institution's repository design methodology.

The growth in any domain of knowledge rarely develops cleanly. I have attempted to present the development of archival thought and the associated repository architectures in as succinct a line of intellectual succession as possible for the sake of clarity, while fully understanding that it's often difficult to acknowledge any single source for new intellectual ideas. Additionally, the research into digital repositories cuts across several domains of knowledge, with little agreement on the semantics of words like *information*, *record*, *item*, *data*, etc. I have appended a glossary that defines particularly fuzzy terms for the reader. These definitions are drawn from differing domains of knowledge in the information science world, and I have noted the source of each.

3: Principles and Practices of Archival Science

Museums, libraries and archives are cultural heritage institutions with overlapping missions, though the archival community's intellectual mission may be the most misunderstood of the three.[8] The archivist's methodological approach has developed through the intersection of high-level *principles* and hands-on *practices*. The principles, developed over centuries,[9] guide the archivist in all their decision-making. Hovering just under these are the modes of practice by which the principles become instantiated. Principles and practices have become irrevocably linked in modern archival work, but it is helpful to differentiate the philosophically-oriented principles from the action-oriented practices.

Principles generally inform practices, but certain archival practices are so firmly entrenched in the totality of archival science that it is difficult to determine where principle ends and practice begins. Archival practice need not be replicated in repository designs as long as the principles behind them remain inviolate. It is desirable, however, to apply appropriate archival practices to the design of digital repositories in a manner consistent with their current use. This application leverages the benefits of existing archival knowledge and provides a continuum of experience during the early stages of digital repository development when a stable set of both principles and practices helps to provide a baseline for experimentation.

Within the archival community there are subtle differences in the application of principles and practices across categories of archival materials. The categories are generally identified as archives, personal papers and manuscripts, with archives defined as "the noncurrent records of an organization or institution preserved because of their continuing value,"[10] and personal papers and manuscripts as "a natural accumulation of documents created or accumulated by an individual or family belonging to him or her and subject to his or her disposition."[11] The significant difference between these materials is the method of their agglomeration, and each category of materials includes the complete range of archival media: business records, letters, photographs, sound and video recordings and other ephemera. There are such commonalities amongst archives, personal papers and manuscripts that it is convenient to identify standard archival principles and practices that apply equally to all. [12]

Due to the intrinsically unique nature of individual archival materials, archival scientists have historically resisted efforts at standardization. As James O'Toole noted:

> Every archives collection was different, and therefore it seemed necessary that every archives repository be different, with its own way of doing things. Archivists acknowledged certain general theoretical principles, to be sure, but they were always prepared to allow each repository the freedom to apply those principles in its own way.[13]

Gradually, with an acceleration over the last thirty years, these dispersed practices have become increasingly homogeneous as a result of technological advancements that enable archival information to be accessed in electronic networked environments. The appearance of Steve Henson's *Archives, Personal Papers and Manuscripts* in 1983 attempted to reconcile archival cataloging and description with the conventions of 1978's *Anglo-American Cataloguing Rules, 2nd edition* to provide a basis for the incorporation of archive and manuscript records into online catalogue systems.[14] This was followed by the establishment in the early 1980's of MARC AMC, the archival world's version of the machine-readable cataloguing record. The explosion of interest in the World Wide Web pushed archival scientists to explore ways to make their materials discoverable in the networked environment, which led to recent efforts such as the Dublin Core metadata set for descriptive metadata,[15] and the Encoded Archival Description Document Type Definition (DTD) for archival finding aids. [16]

These standards-making efforts had the effect of enforcing consistency across differing archival repositories, and pushed archival institutions into practical and theoretical concord with each other. The broad acceptance of these standards has consolidated consensus in the archival community around a set of principles and practices that forms the essential core of archival science.

3.1: Archival Principles

The turn of the century has brought a welcome reexamination of archival principles and practices in the light of the influence of electronic records and the World Wide Web on the archival weltanschauung, but consensus remains on a core set of principles that guide the tenants of archival science:[17]

- The Sanctity of Evidence
- The Preservation Imperative
- The Primacy of the Record
- Respect des fonds, Original Order, and Provenance
- Hierarchy in Records, and their Collective Descriptions

There is no strict division between these conceptual areas, and each are inherently interrelated to the others. The conceptual development of *evidential value* naturally flows into a consideration of the archivist's traditional custodial role, which implies agreement on the essential components of archival materials, which directs archivists to devise common approaches to appraisal and description.

3.1.1: The Sanctity of Evidence

"Evidence in the archival sense can be defined as the passive ability of documents and objects and their associated contexts to provide insight into the processes, activities, and events that led to their creation for legal, historical, archaeological, and other purposes."[18] This conception of the evidentiary quality of archival materials has a long history in archival science, suggested by Jenkinson in 1922's *Manual of Archive Administration*, though only in passing at that time.[19]

T. R. Schellenberg, an influential theorist and longtime employee of NARA, developed a more refined concept of evidentiary value in writings discussing the appraisal methods of public records.[20] He described public records as having two distinctly different values, *evidentiary* and *informational*, both of which might be found in a single record. *Informational* value in records was the information they contain about persons, corporate bodies, things, problems, conditions, etc., with which the Government body dealt, while *evidentiary* value was defined as the evidence documents contained of the organization and functioning of the Government body that produced them. The knowledge derived from Schellenberg's evidentiary concept is clearly focused on the creating organization, but he later tempered the organizational centricity of evidential value through his additional contribution of the concept of the *primary* and *secondary* values of archival materials. These conceptions expanded the definition of evidentiary value to include both the evidence of organizations and the informational value contained in their records. This widened conception of evidentiary value was recognized by Gilliland-Swetland, who noted that "evidential value in the widest sense is reflected to some extent in any information artifact, but only a subset of all information is subject to legal or regulatory requirements concerning creation and maintenance."

There is an identifiable evidentiary essence in all potential archival materials, and it is a significant duty of the archivist to preserve the values imbued in this essence. A consideration of the sanctity of evidence leads directly to a consideration of the preservation imperative, a critical area of consideration when confronting digital materials, for it is during the process of attempting to preserve digital materials that evidential value is most often at risk of being compromised.[21]

3.1.2: The Preservation Imperative

> It has long been the responsibility of libraries and archives to assemble, organize, and protect documentation of human activity. The ethic of preservation as coordinated, conscious management, however, is a more recent phenomenon. Librarians and archivists—like the clerks and scribes who went before them—have increased the chances that evidence about how we live, how we think, and what we have accomplished will be preserved. Traditional preservation, as responsible custody, works only when this evidence has a physical form; when the value of the evidence exceeds the cost of keeping it; and when the roles of evidence creators, evidence keepers and evidence users are mutually reinforcing.[22]

The conception of preservation as coordinated, conscious management arises from Jenkinson, and his statement that "archive quality is dependent upon the possibility of proving an unblemished line of responsible custodians."[23] The concepts of stewardship and responsible custody have been the theoretical constructs supporting archivists as their role has changed from conservators to preservationists. This becomes especially true as the archivist moves from paper materials to electronic ones. As Trudy Peterson stated eloquently in 1984, paper records force the archivist to confront issues of storage, while electronic records force them to confront preservation. [24]

Traditional archival custody began the moment materials were accessioned into the archive. For electronic materials, the influence of the information life cycle concept, rising from the records management domain, has expanded the conception of when materials come under consideration for archival care. The traditional information life cycle has five stages: the creation stage; the distribution and use stage; the storage and maintenance stage; the retention and disposition stage; and the archival preservation stage.[25] The archival community is increasingly involved at points further upstream in the life cycle of records, working closely with records managers to influence information management decisions as early as the creation stage. But while the introduction of the life cycle concept has extended the temporal range of archival control, the custodial and responsibilities have remained the same. These custodial responsibilities include administrative responsibility, organizational viability, financial sustainability, security and procedural accountability.[26]

3.1.3: The Primacy of the Record

The item commonly identified as the smallest granular unit in digital information systems is the *bit*. Preservation at the bit level is an important goal of digital repository systems, but the concept of the *bit* fails to accurately describe the complex nature of archival materials. Bits become more interesting when they are ordered into higher level entities (bytes, sets, bags, information, knowledge, content, digital objects, records) just as phonemes are more interesting when they're ordered into novels, contracts or correspondence.

Archivists have gradually come to coalesce around the concept of the *record* as the baseline component of archival descriptive practice. Archives and historical manuscripts have traditionally been defined as "the records, in any physical form, produced by organizations or individuals in the course of activity over time, and then saved permanently for some further use,"[27] while particular definitions of the word *record* have been elevated by different branches of archival practice. A record is not the most granular item found in an archival system,[28] but the concept of the record fits comfortably into traditional archival system of descriptive practice, and the structure of records corresponds favorably to the complex digital objects found in electronic information systems. Richard Cox's definition, originating in the records management domain, emphasizes the complex nature of both traditional and electronic records, and identifies three significant components:

> We have long possessed a working definition of records, stressing the fact that they document a specific activity or transaction and that this documentation has a particular content (information), structure (form), and context (relationship to a creator, function, and other records). A record is a specific entity. Records are transaction oriented. They are evidence of activity (transaction), and that evidence can only be preserved if we maintain content, structure, context. Structure is the

> record form. Context is the linkage of one record to other records and to the originating process. Content is the data or information, but content without structure and context cannot be data or information that is reliable.[29]

The constancy of archival materials has also been a subject of considerable debate. The previously held judgment of *permanent* value as a criterion for preservation has increasingly come into question,[30] with the current warrant applying the application of *enduring* value. "Enduring value stems from a document or record's intrinsic attributes, the contextual documentation that surrounds it, its relationship to other records and entities, and assurance of its authenticity and reliability."[31] The move to the concept of endurance as opposed to permanence gives the archivist decision-making flexibility, but O'Toole has argued that if the conception of permanence is to shift in any way, it should shift in terms of refocusing attention on the permanence of the information *in* records rather than on the items themselves.[32]

Traditional archival theorists have made attempts to define the central attributes of archival records (Jenkinson chose *impartiality* and *authenticity*),[33] but a concentrated focus on the attributes of records has only recently garnered widespread attention with the consideration of the essential components of electronic records. While Cox' Pittsburgh Project considered the functional requirements for recordkeeping systems, concurrent activities approached records from other angles, extending archival thinking further into the consideration of digital materials.

Two years prior to the completion of the Pittsburgh Project, Luciana Duranti of the University of British Columbia wrote presciently on the concepts of reliability and authenticity in records, introducing scholarship that supported the evidentiary value of archival records. In her view, the evidential value of a record is based on its reliability;

when it can be treated as a fact in itself, as the entity of which it is evidence. A record is authentic when it is the document that it claims to be.[34] Reliability and authenticity come together to determine the *integrity*, or genuineness, of records:

> Genuineness is the closest concept to truthfulness. It is generally accepted by all literate civilizations that documents that are trustworthy (that is, *reliable*) because of their completeness and controlled procedure of creation, and which are guaranteed to be intact and what they purport to be (that is, *authentic*) by controlled procedures of transmission and preservation, can be *presumed* to be truthful (that is, *genuine*) as to their content. Thus, to those who make and preserve records, the two key concepts remain reliability and authenticity, as genuineness is embedded in them. [Author's italics].[35]

Duranti's views on the authenticity and reliability of records arise from her interest in the study of diplomatics, a seventeenth century analytical technique for determining the authenticity of records issued by sovereign authorities,[36] but her approach is especially useful when considering digital information.[37] The Interpares Project, of which Duranti is a task force member, has published a task force report on authenticity in electronic records that provides a detailed method for ensuring that elements of authenticity and reliability are retained in electronic information systems.[38]

Duranti's focus on the *integrity* attribute of records was echoed in the 1996 Research Libraries Group (RLG) report "Preserving Digital Information," informally known as the *Waters/Garrett report* after its co-chairs, John Garrett (then of the CyberVillages Corporation, a subsidiary of the CMGI corporation), and Donald Waters (then the associate university librarian at Yale University, and now at the Andrew W. Mellon Foundation). The report remains the most significant document bridging traditional and digital archival concepts, and in particular, provided the most complete and successful model to date of the essential attributes that should comprise an archival digital object. The report finds sympathy with the archival conception of the *record*, but chooses the term *digital object* in its stead when discussing record-like items in digital systems. I follow this convention.[39]

The section of their report entitled *Information Objects in the Digital Landscape* focuses on the issues of representation and integrity of records over the long-term, and identifies five core attributes (*conten*t, *fixity*, *reference*, *provenance* and *context*) that must be represented in an accurate expression of a digital object.[40] *Provenance* is such a central tenant of archival practice that I deal with it separately. *Content* and *context* have been touched upon in the earlier definition of the archival *record*. *Reference* asserts that the object must always be *findable,* with an orderly discovery system in place. That leaves the more abstract term *fixity*.

David Levy has referred to records as talking things, implying that they are bits of the material world that their creators have imbued with the ability to speak as surrogates on their behalf. The most important thing that archival records can do is to hold talk *fixed* to ensure their repeatability.[41] A consideration of fixity has long been important in physical archives (the conservation arts are largely tasked with the retention of information fixity in traditional forms), but it is in the digital realm that the issue of fixity becomes most urgent.

As Garrett and Waters noted, "if an object is not fixed, and the content is subject to change or withdrawal without notice, then its integrity may be compromised and its value as a cultural record would be severely diminished."[42] The malleability of digital information has always been one of its most beneficial qualities, but the ease of alteration also makes digital information inherently unstable. The ability to secure the fixity of information in digital form directly relates to the degree of integrity that can be engendered in electronic systems.

3.1.4: Respect des fonds, Original Order, and Provenance

When archivists look at records, what they see first and foremost are provenance and original order. This perception permits the establishment of a context that will serve as the basis for everything else the archivist does. *Provenance* is the fundamental principle of modern archival practice, and refers to the method of preserving the organizational context and course of activity in a set of records which is captured in the materials in their original state.[43] The modern invocation of provenance arises out of two European traditions; the French notion of *respect des fonds*, and the German concept of *original order*, and is essentially a conceptual join of these related ideas. Respect des fonds relates to the disposition of materials by organization, and implies that records should be grouped according to the nature of the institution that accumulated them.[44] Original order entails that records are "maintained in the order and with the designations which they received in the course of the official activity of the agency concerned."[45] An arrangement of materials based on the principle of provenance implies that the original filing order and overall structure of a collection "is not the result of chance, but the logical consequence of the organization of the administrative body of whose function the archival collection is the product."[46] Provenance in the consideration of digital materials inherited additional aspects from the museum, art and architecture communities.[47] The definition of provenance in these communities is similar to the archival definition, but places greater emphasis on the successive transfers of ownership and custody of a particular manuscript,[48] as opposed to the archival focus on the contextual organization of materials in their original state.

The archival sense of provenance implies that electronic records be accessioned into digital repositories with an ordering by institution, and that they remain (as much as is possible when considering computer file systems), in the original order in which they were created. At the same time, the custodial definition of provenance manifests a process that ensures the integrity of records in digital environments. Electronic records have an inherent mutability, but their authenticity, reliability and integrity can be assured through a custodial monitoring of the transfer of ownership of digital objects within a system, and by noting any changes that might take place in the files through migration, refreshment or unforeseen information loss. In digital repositories, the principle of provenance is achieved through the use of software components that manage audit trails, versioning and workflows.

3.1.5: Hierarchy in Records and their Collective Description

Schellenberg conceived that the file structure of records revealed information about the functions and activities of an organization, and thus the evidential value of records. The records of individual organization arrive at an archive in a naturally hierarchical structure, and the archivist must maintain that structure to retain the evidential value embodied in the records of that organization. Hierarchical structure in archival records is guided by the concept of *levels of control*.

Levels of control explain the choices made by the archivist while implementing the concepts of provenance, respect des fonds and original order:

> Most modern archival work involves progressively grouping and describing sets of records along a continuum from the largest and most general to the smallest and most specific. Thus the records of an agency can be successively both physically subdivided and intellectually described in terms of its constituent offices, activities, or functions; the files within each series; and the documents within each file. Each of these refinements is regarded as a different level of control.[49]

Levels of control are imposed in electronic file systems by directories and folders, but the control decisions are much more subjective in traditional archival practice. The principle of levels of control applied in the electronic environment allows granularity down to the lowest level necessary, but also assures that items can be physically and intellectually organized in successively larger groups based on their provenance and original order requirements.

The archivist describes these hierarchical levels collectively, rather than as discrete units. "In contrast to artificial groupings, the documents in archival collections relate to each other in ways that transcend the information in each document. The archival whole is greater than the sum of its parts; the relationships are as important as the particulars."[50] Exceptional archival objects may be considered individually, but archival items are generally grouped in collections. Collective description has developed out of the necessity to provide an overall view of a collection without the benefits of unlimited resources for item-level cataloging. Collective description also acknowledges the value a higher-level view of a collection can provide to researchers.

3.2: Archival Practices

The archivist implements the principles above through practice. Implementation begins at appraisal, and continues through the arrangement and description of the materials. The archivist's efforts to *appraise*, *accession*, *arrange* and *describe* are ultimately at the service of their mission to provide access to the materials. The principles of the sanctity of evidence, the focus on the record as the central archival unit, provenance, hierarchical arrangement and collective description find their application in the tasks of archival practice, and it is at the appraisal point that these concepts are first applied.

3.2.1: Appraisal and Accession

The initial task of the archivist is to examine newly available materials to determine their suitability for addition to the archive. This is the negotiated process of *accession* between the archive and the donor that culminates in an deed of gift, a document codifying the legal rights the archive has to the material with which it is taking possession. The completion of the deed of gift is the last step taken prior to the archive gaining full possession of the materials. "Accessioning comprises all the steps that repositories take to gain initial physical, administrative, legal, and intellectual control over newly acquired material,"[51] and *appraisal* is the set of tools at the archivist's disposal to aid them in making those determinations. Accession and appraisal is a complex process of decision-making involving the application of rules determined by the mission of the archive and its collecting policy, the education and knowledge of the archivist, the access to information about the collection being added, and numerous other factors. As noted earlier, the process of appraisal can begin at any number of points along the lifecycle of records, though the actual possession of the materials by the archive really begins at accession.

Intellectual appraisal begins as the archivist assigns content, context and structural tags to archival items in the accession process. The consideration of Schellenberg's *primary* and *secondary* values is but one of many intellectual decisions made by the archivist at this point. Materials with primary value are self-evidently important to the archival organization, but the secondary value of materials are more subjective. To be considered to have secondary value, the material under consideration should show evidence of the organization of an institution, or evidence of the information about that institution (or both).

In practice, Schellenberg suggested ways to apply tests of evidentiary informational value to the materials being appraised. The tests for evidential value must first be made while considering the entirety of the agency's records, and must not be made on a piecemeal basis. A series of evidential checks are made that attempt to identify the chief functions of an agency, their relative importance, and which units have the primary responsibility for decision-making.[52] Not all agencies pass the importance test, and Schellenberg noted that it was "a curious anomaly that the more important a matter, the less likely is a complete documentation to be found."[53]

The tests for informational value include tests for the *uniqueness* of the information and the uniqueness of the records that contain the information;[54] the *form*, that relates to the degree in which "records that represent concentrations of information are the most suitable for archival preservation, for archival institutions are almost always pressed for space to house records;"[55] and finally, the importance of the information in the records. Archival records must pass the uniqueness and form tests before importance can be considered, because importance can be a relative concept, while uniqueness and form are much more concrete.[56]

Practical appraisal refers to the physical storage methods and materials, costs of storage, estimates of usage, records of item transfers and the like that must be considered before items can be accessioned into an archive. These are the challenging resource decisions that must be made by archives outside of any intellectual decisions they might make about the materials. Conway noted that the essence of preservation management is resource allocation, and that "people, money and materials must be acquired, organized and put to work to ensure that information sources are given adequate protection...Cost-effective preservation action cannot take place without compromising ideal outcomes."[57] Such are the difficult decisions that must be made when undertaking practical appraisal.

3.2.2: Arrangement

The principle of provenance discussed earlier governs entirety of the arrangement decisions. The archivist first considers the body of work as a whole, keeping the principles of respect des fonds and original order clearly in mind. Jenkinson observed that "the only correct basis of arrangement is exposition of the administrative objects which the archives originally served,"[58] additionally noting that effective arrangement should be based a full understanding of an agency's mission garnered through research into the agency. Additionally, while it's easiest to think of arrangement as happening at the beginning of the accessioning operation, it is more accurately depicted as an iterative process that evolves as the archivist learns the materials while dividing it into hierarchical levels.

Oliver W. Holmes codified the arrangement control levels while working at NARA, and published his findings in a widely cited article in the American Archivist.[59] He divided physical control into five levels; repository level; record group; series; file unit; and document [record]. The top level of structure within a repository, the record group, roughly corresponds to the earlier conception of the fonds. These record groups can be further divided into subgroups based on administratively discrete units or activities that produce records.[60] These hierarchical levels are implemented in practices that consider the arrangement of materials by provenance, and the arrangement of materials by filing structure.[61] Arrangement by provenance is an intellectual form of arrangement, but as Miller points out, reality does not necessarily arrange itself that simply:

> Archivists are aware that complex institutional networks and relationships increasingly represent reality better than the simple monohierarchical model in which every office has one superior...Archival arrangement should thus not be thought of as one unified system in which physical files and file series are at some lower level than record groups, collections, and subgroups. These are instead two different kinds of arrangement—arrangement by provenance/records creator and arrangement by filing structure.[62]

3.2.3: Description

Ken Haworth compiled definitions of *archival description* from several sources, the most epigrammatic of which is "the recording in a standardized form of information about the structure, function and content of records."[63] Archives have traditionally utilized the construct of the *finding aid* as the device through which archival materials are publicly described. "The basic National Archives 'finding aid' was not a card catalog, but a pamphlet-like descriptive inventory of each Record Group, containing introductory information about the records as a whole and a list of all the constituent series,"[64] and other archival organizations have followed this pattern. Steve Henson's *APPM* attempted to fit the imprecise structure of the finding aid into the highly standardized structure of a MARC record to enable the information in the finding aids to be added to library public access catalogs. This effort has now largely been supplemented through the use of the Encoded Archival Description (EAD) DTD, which wraps archival finding aids in the Standard Generalized Markup Language (SGML) for distribution over the world wide web. As Ken Haworth noted:

> The plethora of finding aids generated by archivists (MARC collection/item-level catalog records, inventories of fonds/collection-level and series-level descriptions with associated box and files lists) were contained in a variety of software-dependent systems which could not be easily linked or exploited. Clearly there was a pressing need for an Internet-accessible archival description system that integrated and linked all of the descriptive records that make up a fonds/collection description.[65]

In practice, collective description has a number of advantages.[66] First, it documents all the records of the same provenance. Secondly, it permits economies in description, whereby an archivist can describe materials only to the level and detail of their expected use. Third, it mirrors the arrangement of the materials, and fourth, it can be applied regardless of the nature of the materials and does not require specialized description for special forms of materials. Additionally, hierarchical collective description can provide for the inheritance of one level of description by the level below.

3.2.4: Access

Most of the decisions the archivist makes in evaluating principles and practices are made to provide access to the materials under their care. While archivists are often considered guilty of loving the materials more than the users of those materials, without access the archival goods serve little purpose.

Physical archives have composed detailed access policies that state clearly the materials that can be accessed; who can access them; for what purposes; and with what rights attached. Access may also be determined by the physical condition of the materials, or by the hardware necessary to make the materials understandable. Issues of access also directly touch upon issues of security.

The process of providing access in the physical archive is relatively straightforward. An employee of the archive accepts the request of a patron, gathers the material from the vaults, and then distributes them to the patron in the manner accorded by the repository's rules. Some items requested by the patron are restricted and cannot be obtained. Other items are so fragile or valuable that they can only be shared as photocopies or in sound recordings available in access-only copies on compact discs or cassettes. Other items can be viewed by the patron, but he must view them in a restricted area under the watchful eye of the archivist, and is forbidden from copying them.

The challenge is translating these access and security behaviors in the traditional archive into concomitant behaviors in the digital repository.

4: Translating Archival Principles and Practices into Digital Repository Architectures

Despite the accelerating pace of digital creation over the last quarter-century, the public remains blissfully unaware of the hazards that face information caretakers in the near future due to the proliferation of digital information and our current inability to guarantee its viability for anything close to the long-term. Without the development of systems that can address the issues surrounding the preservation of digital materials, the world is in danger of suffering a catastrophic loss. "The archival science perspective can make a major contribution to a new paradigm for the design, management, preservation, and use of digital resources."[67]

Despite its reputation as a haven for the historically-minded, the archival profession has traditionally been very active in exploring information management breakthroughs. At every burst of innovation, archivists were there to ensure that the archival paradigm of an evidence-based approach to information management continued to be articulated.[68] The digital information explosion has created profound preservation dilemmas, and as the archivist brings a wealth of applicable knowledge and practice to the subject, it is only natural that the archival community would become aggressively involved in digital preservation issues.

NARA and the Library of Congress (LC) are the two United States government institutions that have traditionally taken the lead in addressing electronic records issues from the archival perspective, but their engagement with electronic records issues has been a bumpy ride.[69] Still, many other archival organizations look to NARA and LC for leadership, and they have both provided guidance in the development of digital repositories.

Ground zero in the movement to address concerns about the tenuous viability of the digital information was the publication of the *Waters/Garrett report* in 1996. The report proposed a clear research agenda for digital preservation exploration, with specific proposals including the encouragement of studies that would explicitly explore the design of systems to facilitate digital archiving, and systems capable of storing massive quantities of culturally valuable digital information. The exploration of digital repository system design was thus placed centrally in the future research agenda.

The concept of a digital repository is actualized in an electronic information system as a software architecture and its associated management responsibilities that aggregates an array of tools, hardware components and business rules into a coherent system capable of providing the necessitated range of preservation-oriented functions. These repository systems would effectively replicate the functions found in a traditional brick-and-mortar archive to the extent that those functions were necessary in the digital realm. Despite the tremendous elucidation of the most pressing issues that the *Waters/Garrett report* had been able to provide in 1996, the process of articulating the principles and practices of archival science and placing them in a high-level repository design architecture was a slow process.

The research community eventually came to coalesce around the high-level design provided by the *Reference Model for an Open Archival Information System (OAIS),*[70] a dense 150 page document that defines the functional requirements of a physical or digital archival system that purports to address long-term preservation requirements.[71] The process of developing the reference model into an international standard took over seven years. (In the following sections we will refer to the report itself as the *reference model* (RM), and an implementation of the reference model in a system as an OAIS.)

The model was developed by the National Aeronautics and Space Administration (NASA) in collaboration with over two dozen international space agencies under the name of the Consultative Committee for Space Data Systems (CCSDS). While the effort was being driven by space scientists, the archival community had a presence in OAIS development from the first U.S. workshop held in October of 1995.[72] Both Bruce Ambacher, an information technology specialist with the Modern Records Program of NARA, and John Garrett, one of the co-authors of the *Waters/Garrett report* (at the time an employee of Hughes STX), were attendees at the first workshop meeting, and NARA eventually hosted thirteen out of the sixteen U.S. meetings up through 1999, at which time the RM was under formal review by the International Organization for Standardization (ISO).

Concurrently with his participation in the RM meetings, Garrett had been working with the Task Force on Archiving of Digital Information, and he was able to introduce the task force's finished report as reference material to the Fourth U.S. Workshop held July 10-11, 1996. The *Waters/Garrett report* was of essential value to the RM designers, holding its place as one of thirteen informative references that were featured in the final report released in January of 2002. On February 24, 2003, that final recommendation, also known as the Blue Book version of the OAIS, became ISO 14721:2003, an international standard.

The RM has many merits, but still provides only a conceptual framework for repository design, remaining agnostic to possible implementations, and only hinting at possible real-world solutions from its high-level perch. It has gained quite a bit of traction in the research community, though as Mackenzie Smith of M.I.T. has suggested,[73] it's chief value may be the establishment of a neutral vocabulary by which researchers in different knowledge domains can communicate about similar concepts without arguing over semantics.

Despite its difficulties, the RM is the document that translates archival principles and practices into a high-level digital system design. The language of the RM informs all subsequent research initiatives that deal with digital preservation, and especially those that address the design of repositories for long-term preservation.

5: The OAIS Reference Model

> There is a tension in digital preservation research between generalizable principles, methods and technologies that cut across formats, content areas, academic disciplines, and institutional settings, and the very specific requirements of different producers, content types and user communities. Organizations facing immediate and pragmatic concerns usually will do enough research to meet their current business needs, but typically will not be motivated to analyze their work for its wider relevance or applicability.[74]

One of the significant obstacles archivists have confronted when attempting to evaluate digital repository architectures is the difficulty in identifying a common vocabulary between traditional archival practice and the language of integrated information management systems. Electronic information systems incorporate data management concepts similar to those found in archival (especially records management) practice, but the terminology has subtle differences, and it has been a non-trivial matter to uncover the equivalencies between the glossaries of the two fields.

The RM solves this problem. At first blush it has the misfortune to appear both highly technical yet vaporish. As the reader's familiarity with it increases, however, it unfolds into a prescient document that creates a common language allowing existing and future archival systems to be compared and contrasted. The applicability of the RM to both traditional and digital archival entities signifies how deeply the authors incorporated traditional archival thinking into their model, and allows a direct mapping between the RM's more complex concepts and terminology familiar to archivists. The RM terminology, fully vested with archival meaning, has become the common currency of digital repository designers, and the research efforts enumerated later in this paper describe their efforts using the terminology found in the RM.

The RM is specifically designed as a requirements model for any archive, physical or digital, that seeks to provide for the permanent, or indefinite long-term, preservation of information, but has become an especially popular guide for the design of digital archiving systems. The RM defines the concept of *indefinite long-term* as a time period long enough to be concerned with the impacts of changing technologies. Paul Conway has suggested that the same concept be applied on a generational basis, that is, that the indefinite long-term be defined as the point when items can be passed on to the next generation of overseers.[75] A wit has noted that "Digital documents last forever—or five years, whichever comes first."[76] Though there is still disagreement on what exactly *long-term* might imply, the need for extended stewardship has already been established.

The attributes that invoke the perception of the RM as a difficult work also provide its benefits. The terseness that diminishes its pleasure-reading value enables it to make its complex points with clarity, and the repository system model it describes unfolds with precision and comprehensibility. It addresses the major activities of an information-preserving archive in order to define a consistent and useful set of terms and concepts, understandable across divergent domains of knowledge.

The acceptance of the RM as an international standard and its relatively rapid adoption in both technical and archival communities facilitates developments in the design of repository architectures that converge along similar paths. The adoption of the RM among designers provides a standardization of components and functional areas, and should promote greater vendor awareness of archival needs in the commercial sector. This last development should encourage the commercial development of preservation systems, or the active inclusion of archival functionalities in commercial Content Management Systems (CMS),[77] though a strong commercial presence in the design of preservation-oriented repositories has not yet been seen.

An OAIS has both horizontal and a vertical components. The horizontal aspect is represented in the RM by the *Functional Model* of the system, which refers to the workflow components of the system. The *Functional Model* tracks the six different functional entities of an OAIS and their related interfaces as they move across time from the moment a complex digital object enters the system, through the time it is managed by the system, and on to a point of dissemination at the other end. Each of these functional entities are high-level conceptual models of a set of behaviors surrounding a set of requirements, and as such can include both human and system hardware/software components. The six entities are *Ingest*; *Preservation Planning*; *Data Management*; *Archival Storage*; *Administration*; and *Access*.

The vertical component is represented by the *Information Model*. The *Information Model* approaches each complex digital object individually and examines it in great detail, describing the layers of content, fixity, reference, provenance and context metadata that form the digital object and that enable it to be self-validating and self-instantiating.[78] The RM uses its own complex terminology to describe these five different layers of information.

The vertical and horizontal models are intricately related. The *Information Model* describes the nature of the complex digital object, and the *Functional Model* describes the entities that interact with the digital object within the system. In addition to the *Information Model* and the *Functional Model*, the RM includes information about how digital object producers, consumers and management personnel interact with the system, even as it models these entities as standing outside of the scope of an OAIS system. While most OAIS repositories are determined to be primarily the receptacles of deposited items, the RM also mentions a particular type of archive, known as an *active archive*, in which the producer role and the archive role are the responsibility of the same entity.[79] Active archives have functionalities similar to some CMS.

Another important concept in the RM is a special category of consumers known as the *Designated Community*. A consumer in the RM is the role played by people, or other information systems, that interact with the OAIS to find and acquire preserved information of interest. The *Designated Community* is the set of consumers who should be able to understand the information housed in a particular OAIS.[80] The concept of a *Designated Community* comes up often in the RM when referring to forces that can influence changes within a particular OAIS, and the term is a useful one to describe the residual influences that information systems and their community of users have on each other. The RM also discusses functionalities that it refers to as *common services*, including an operating system platform, network connectivity and common security services, that fall outside of the scope of the RM, but are essential for the operation of any digital repository.[81] The choice of operating system may be a preservation decision as much as it is a business decision.

The review of the *Functional Model* and the *Information Model* that follows provides an overview into the broad capabilities of an OAIS, identifies the genesis of OAIS terminology in the previously defined archival principles and practices, and notes specific software instantiations that might fall under each broad functional area. This set of functional requirements is designed as a list of computer architecture components, but the analysis of these components will also lead us into discussions concerning the roles people play in preservation systems and activities.[82]

5.1: The OAIS Functional Model

See Appendix A for a detailed view of the Functional Model.

The Functional Model divides the OAIS into six different functional entities: *Ingest*; *Archival Storage*; *Data Management*; *Administration*; *Preservation Planning*; and *Access*. Each of these functional areas is an agglomeration of various rules and work processes, but the RM language describing these processes can be quite confusing. The easiest way to grasp these abstractions is to imagine each RM concept as relating to a task that a human archivist might perform in a physical archive. One example is exemplified by a look at the *Error Checking* sub-function of the *Archival Storage* functional entity. The RM description of the *Error Checking* sub-function is somewhat confounding: "The Error Checking function provides statistically acceptable assurance that no components of the AIP [archival information package] are corrupted during any internal Archival Storage data transfer."[83] This sounds quite complex, but the human analogue allows us to see the high-level function in much simpler terms. Imagine, if you will, the following scenario: an archivist examines (error checking) materials (AIPs) returned by a patron (internal archival storage data transfer) to assure (statistically acceptable assurance) that they are in the same condition and order they were in as when they were first given to the patron (no components...are corrupted). The RM is quite successful in mapping the traditional archival duties to high-level functions that can be applied to computer systems, but the complexity of the language they use can sometimes get in the way. Translating the RM terminology into archive-ese is one goal here.

Each of the six different functional entities mentioned above are essential to the overall operation of a successful OAIS, but not all of the entities are essential to core long-term preservation functions. As we review the entities we will attempt to extract a smaller subset of functions highlighted as core preservation functionalities essential to actual system implementations.

5.1.1: Ingest

The choice of the term Ingest has been met with disapproval in some quarters due to its unfortunate connotations of food intake. Though it does evoke some unpleasantries, the term is agnostic to origins and data types of the material being entered into the system, whereas concepts like *acquisition*, *authoring*, or *accession* are loaded with semantic meaning drawn from particular domains of knowledge, and that prevent them from being as widely applicable.

The Ingest entity includes all the processes necessary to prepare a digital item for entry into the archive. In archival repositories most of the authorship takes place outside of the scope of the archive. An OAIS repository generally acquires the material that it stores from other locations, though this is not necessarily the case. The custodial responsibilities of the archivist begin at the moment of submission, and these responsibilities, which include procedural accountability and administrative responsibility for the materials under their care, are of exceptional importance when considering digital repository architectures.

The ingested digital objects can appear in a variety of forms; plain text documents, sound files, graphics, or a complex multimedia combination of any of these (as in electronic journals). The record, in archival practice, is a complex item that must be described in terms of its informational content, its physical structure, and the context in which it is situated. The digital object in a repository is also complex, a creation of bits and metadata, and for it to retain its archival values, the digital object must be able to embody all the characteristics of the archival record. In practical consideration, this is done by surrounding the core of the digital object (the *Data Object*) with layers of *Representation Information*, that is, metadata.

As an army runs on its stomach, and a library runs on its acronyms, so must a digital repository run on its metadata. In order to determine whether a digital item has retained its intellectual content from the beginning of its life to a later portion of its life, the item needs to be fully defined at the beginning of its life, in terms of its configuration of bits, the structure and format of its representation and on the ideas that it contains.[84] Metadata is the key, however, the system needs to be able to ingest and manage the bitstreams and the associated metadata separately from each other.[85]

The high-level concepts of provenance, levels of control, and collective description form the cornerstones of archival practice, and they must be applied to digital items at their moment of submission to the archive. This implies that the traditional archival processes of selection have been applied to items prior to the decision to ingest them into the system, and also that the negotiated agreements between content suppliers and the OAIS must be in place so that provenance, levels of control and collective description can be implemented on digital items at the earliest possible time. As the digital objects move through the workflow diagrammed in the *Functional Model*, they are catalogued and documented, and metadata is appended to them.

A confusing RM construct are the different names given to the same digital object as it moves through different phases of its lifecycle. The RM assigns digital objects into three different packages, depending on their position in the workflow, and the item's position in the workflow determines its requirements. The three phases are delimited from each other because at each phase of the workflow the digital item can have different metadata, rights, permissions and representations associated with it. The *Submission Information Package* (SIP) refers to the digital object and its associated bundle of metadata at the moment of submission, with the information largely provided by the author; the *Archival Information Package* (AIP) is the same digital object after archive-dependent metadata has been added; and the *Dissemination Information Package* (DIP) is the digital object plus its associated metadata that has been prepared in a form suitable for dissemination from the archive. The structure of each of these packages is discussed in more detail in the *Information Model*.

Acquired materials often come with some metadata already attached to them, and an OAIS will negotiate with submitting organizations to receive the digital objects in a particular form (perhaps a set of prearranged file formats) and with a particular set of metadata attached (Dublin Core or MARC data, for example). The quality and form of metadata of these acquired objects is largely out of the OAIS control, and the OAIS may have to make changes to the objects once they are submitted into the archive to bring them under intellectual control.

One of the significant potential benefits of digital repository systems and automatically generated metadata is that the deposited content can be catalogued to some extent at the item level, and this granular description provides for the flexibility and reuse of materials. The principle of collective description can be applied by archivists in electronic systems (EAD finding aids, for example), but increasingly, administrative and structural metadata about individual items will be automatically extracted when the items are created or accessioned into the digital repository. The system should be capable of automated error checking, both to determine whether the object has been entered into the system in the correct form, or whether the object already exists in the system. Processes for batch loading should also be available.

These ingest systems can be as simple or as complicated as the designers wish them to be, ranging from the Unix/Linux command line tools for uploading files, through off-the-shelf ftp clients, to elaborately designed graphical interfaces (GUIs). There is no standard at this point, though ease of use will increasingly come into play as the users of the system move beyond the designers and into the libraries and archives. The questions of authorization also arise at the ingestion stage.

This process by which an OAIS negotiates with content submitters corresponds to the movement of archival intervention upstream in the information life cycle model. The Harvard e-journal archiving project[86] provides an example of a submission negotiation process. E-journals can conceivably appear in an infinite number of formats and with any amount of metadata attached (including almost none). The Harvard librarians recognized that it is more difficult for the archive to make metadata additions or corrections in the process of preparing SIPs to become AIPs that it would be for the publishers to provide the information in a standardized form upon submission, so they pushed the responsibility for metadata standardization onto the publishers as part of the submission negotiation. Additionally, the Harvard archive would accept only a small set of preferred normative formats of the digital material.[87] If the publisher did not submit their materials with all of the required information attached then their submission was returned to them. The architecture of the Harvard e-journal OAIS was structured in such a way that it could handle the return of materials until they had been submitted in a form acceptable to the archive.

The Harvard example exemplifies the way the Ingest entity tackles quality assurance issues, just one aspect of its procedural accountability. The digital submission can be validated for quality utilizing a mechanism such as a cyclic redundancy check[88] or a checksum[89] to verify that the item is what it purports to be upon its entry into the archive. These types of validation processes help assure the authenticity and reliability of submitted items, which hearkens back to the earlier discussion of the components of record-ness, and the importance of preserving these elements for evidentiary purposes.

The final step in an OAIS Ingest process is to confirm that the submitted digital item satisfies all of the criteria for membership in the archive. The process of criteria satisfaction represents the movement of the digital object on the workflow from a SIP to an AIP, just as the process of accession moves items from raw and random objects to those that the physical archive has applied some measure of physical and intellectual control. The Ingest entity's duties ends when the final AIP has been created out of the originally submitted SIP.

The process of ingestion in an OAIS is commensurate to the process of accession in a brick-and-mortar archive. Most items in a brick-and-mortar archive are acquired by the archive, though some items, such as finding aids or deeds of gift, are authored from within the archive. Just as a physical archive has procedures for assuring that the archivist is able to gain physical and intellectual control over the items, the OAIS needs to have accession and appraisal procedures that will assist the digital archivist in identifying the evidential and informational values of the materials being ingested in digital repositories that will require preservation.

5.1.2: Archival Storage

The *Archival Storage* entity provides for the permanent storage, management and protection of AIPs, and as such, it's the functional area that most closely reflects the preservation behaviors that take place in physical archives. It is where basic bit storage issues come into play. This includes the structure of the file system, and also includes the amount of storage available, the type of storage (disc storage, near-line, off-line, etc.), and other issues related to the physical management of storage. The repository hardware should be ample enough and flexible enough to allow the collection to grow. Some of the issues relating to how the system accesses the storage structures are considered more fully in the *Data Management* section.

It is important to note that the *Archival Storage* layer has very little knowledge built into it. This knowledge takes place at the interface of *Archival Storage* and various services, which go by the generic term of Application Programming Interfaces (APIs).

The first function of *Archival Storage* is to receive the AIP from the Ingest entity and to place it in the permanent storage facility. The *Archival Storage* entity overseas the management of this storage, including monitoring statistics of use and error logs to ensure all necessary levels of protection for the archived items. The assignment of persistent URLs would take place during the Ingest process, but Archival Storage would be responsible for ensuring that the items remained findable and that the URLs actually did remain persistent. Periodic automated integrity checking of the system and the items that it holds would take place here. The system should be able to detect some minor bit errors as well.

Despite the nascent movement towards open standards in the computer industry, rapid technological change and entrenched proprietary software formats guarantee that current stored data will need to be reconfigured at some point in the near future, and any consideration of the implications of data storage beyond the immediate short term precludes that migration functionalities must be included in the system.

The concept of migration has developed in digital systems to account for rapidly changing software and hardware environments, and the need to keep archived objects understandable under these changing conditions. It is now assumed that digital objects will have to be periodically migrated from one software or hardware environment to another to ensure that the materials are available in an environment accessible to current users (we discuss one possible alternative to migration, *emulation*, in our discussion of the *Preservation Planning* entity below).

The *Archival Storage* entity is responsible for these types of migrations, and may perform any refreshment, replication or repackaging functions that do not cause information loss. An example would be the migration to a new file system where the digital object retained its independent ability to self-instantiate. Any more profound or complex *transformation* processes would come under the authority of the *Administration* entity. The migration functionality is represented in *Archival Storage* by the *replace media* function. The error checking functionality mentioned earlier in this paper is also situated within the *Archival Storage* entity, and supports the efforts to successfully migrate materials without information loss.

Migration is one of the areas yet to be explored in any great detail. The SDSCC, working with NARA materials, has made some early forays, and the Harvard libraries are expected to attempt a migration in the near future,[90] but there is quite a bit of research still to be done in this important area. Mackenzie Smith has taken great pains to note that preservation is an act of human stewardship, and that automated processes only operate at the service of human decision-making.[91] Information refreshment or migration cannot be completely automated because human stewardship and appraisal decisions are an important part of the process.

Additional tools to facilitate migration would include format registries as well as the support for the widest variety of format management tools. Much standards-making activity is taking place surrounding the development of format registries. Format registries assure that knowledge gathered about representation formats is available to systems in the networked environment, and sustainable over the long-term. Stephen Abrams of Harvard is leading the strategic planning behind a registry based in the U.S.,[92] while the UK Public Records office has their own format registry development known as PRONOM.[93] One of the significant points made by the champions of format registries is that the current Multipurpose Internet Mail Extensions (MIME) types,[94] which are the most common public categorization scheme of media types, are not granular enough to capture all of the format information that a repository might wish to capture. For example, the MIME type for a image format such as a TIFF would be "image/tiff." However, there are multiple versions of TIFF files, a condition that could not be represented currently using MIME types as the method of categorization. These efforts are still in the early stages of development.

The alternative to having detailed format information on an unlimited set of formats is to limit the number and types of formats that the system will support to a small, normative set. This implies the development of a set of open standards for basic software types (word processing, spreadsheets, graphics, etc.), which to date has not occurred. Some discussion of this issue will take place in *Preservation Planning*. It is important to note that while there are an almost unlimited number of file formats floating around, there is some consensus on text and image formats that have the greatest uptake. Those text (ASCII, Unicode, etc.) and image (TIFF, etc.) formats are being used by some repositories as their baseline for materials in the repository that can expect long-term support.

The final responsibility of the Archival Storage function is the *provide data* function, which provides copies of the stored AIPs to the *Access* entity and also oversees data transfer operations.

5.1.3: Data Management

There are ways in which the archival concept of *collective description* can be applied successfully in electronic systems. Higher level descriptions should cascade down to lower levels of control in electronic systems, with the previously mentioned EAD document providing an example of how parent/child relationships can exist within documents and collections. This cascade of information, in which higher level description is automatically applied to lower levels of items, is one of the tremendous benefits of database structures present in most digital repository architectures. The *Data Management* entity operates in parallel to the *Archival Storage* entity, and can be most succinctly defined as the database backend for an OAIS:

> An application designer needs organizing principles to handle the large data sets that must be coordinated across a collection of programs. The application designer confronts numerous problems, including data volumes beyond memory capacity, slow performance from unstructured searches, special-purpose code customized for certain storage devices, and varying mechanisms for representing relationships among data elements. These problems arise from computer involvement, not from the intricacies of the application...A database management system (DBMS) provides the needed organizational approach to flexible storage and retrieval of large amounts of data...With simple commands, the application dispatches items to be stored in the database, leaving it to the DBMS to find proper physical storage, to guarantee future access through a variety of pathways, and to integrate the data into meaningful relationships with other elements.[95]

An OAIS should be agnostic in terms of the brand of database system required, but any database system utilized by the OAIS must be able to perform typical database functions such as creating schema and table definitions; creating, maintaining, updating, and providing for the integrity of the data that it houses; for performing queries on the data and for generating reports about the data.

While the functions of the *Data Management* entity seem self-evident to those familiar with database applications, their importance cannot be underestimated. An OAIS requires layers of rich metadata to assure the understandability of digital objects over the long-term, and a rich and flexible database system component is essential for ensuring that those complex metadata architectures can be implemented and maintained.

A digital repository architecture that supports collective description functions should have the flexibility to allow the ordering of collections into hierarchical groups, and then allow descriptive metadata to cascade down through a selection of items if desired. Most digital repository architectures support this type of descriptive practice through the richness of their relational database architectures and their use of XML for encoding items.

5.1.4: Administration

The *Administration* function is fairly self-explanatory and easily understood in comparison to a physical archive. All archival organizations have managing agents that establish standards and policies, control the physical access to the premises, provide customer service, negotiate with donors for items to be submitted, and other management functions. These housekeeping functions are essential to the operation of any archival entity, and these are the duties that the *Administration* entity handles in an OAIS.

However, the *Administration* entity is not only a housekeeping entity: it has direct contact and oversight over all the other functional entities. It is responsible for negotiating the submission agreements, managing the scheduling of previously negotiated agreements, and for confirming that SIPs, AIPs and DIPs all conform to the agreements. It also has responsibility for providing systems engineering to continuously monitor the functionality of the entire archive system and to systematically control changes to the configurations.[96] This includes mechanisms that update the content of the archive. These mechanisms would include the *transformation* decisions referred to earlier.

Additionally, systems that support provenance will support the gathering of information related to successive transfers of ownership and custody of a particular digital object. The retention of workflow information (audit trails) provides integrity checks for the digital objects as they move through the OAIS. The concept of *versioning* (or version control) is analogous to the concept of provenance, in that each successive version of a digital object can be tracked, and objects can be reverted to earlier versions if necessary.

> Information objects in digital form, like those in other forms, move through life cycles. They are created, edited, described and indexed, disseminated, acquired, used, annotated, revised, re-created, modified and retained for future use or destroyed by a complex, interwoven community of creators and other owners, disseminators, value-added services, and institutional and individual users...How reliable the archival process proves to be in the emerging digital environment hinges on the trustworthy operation of digital archives and on their ability to maintain the integrity of the objects they are charged to preserve...For digital objects, no less than for objects of other kinds, knowing *how* to operationally preserve them depends, at least in part, on being able to discriminate the essential features of *what* needs to be preserved.[97]

At the implementation level, the capacity of a system to retain the fixity of content is determined by the business rules of the system (including the read/write/execute policies for particular pieces of content), the levels of security surrounding the system, the ability to back-up the information, and the overall ability to identify and protect a digital object's significant properties. Methods of securing the fixity of digital objects and tracking changes to objects that have appropriate permissions, is one of the keys to building trust in digital systems. Administration is responsible for these considerations.

Finally, it has the *customer service* function, which creates, maintains and deletes customer accounts, as well as providing billing services.

5.1.5: Preservation Planning

The *Preservation Planning* entity operates under the *Administration* entity, but it exercises administration-like control over all aspects of preservation decision-making rather than general management functions. *Preservation Planning* is tasked to monitor the changing technological conditions that would affect the producers and consumers of the OAIS (its *Designated Community*). Changes affecting the *Designated Community* might include changes in data formats, media choices, preferences for software packages, new computing platforms and mechanisms for communicating with the archive. It would also take a forward-thinking approach and track emerging digital technologies, information standards and computing platforms to identify technologies that could cause obsolescence in the archive's computing environment and prevent access to some of the archive's current holdings.[98]

Preservation Planning should undertake periodic analysis of the state of the universe in terms of the development of open standards in file formats and operating systems. The software architecture of the digital repository system should be built on open standards, such as the structured query database language (SQL); the extensible markup language (XML) and other standards that provide for easy interoperability between disparate systems. The openness of the software components frees the architects of the systems from getting locked-in to proprietary software architectures that might impinge on future open development.

Preservation Planning develops strategies and standards for the implementation of preservation functions in the OAIS. This represents itself in the development of SIP and AIP package designs, and in migration plans for AIPs that would prevent the loss of access to those materials due to technology obsolescence. As noted in the *Waters/Garrett report*, "methods for migrating digital information in relatively simple files of data are quite well established, but the preservation community is only beginning to address migration of more complex digital objects...Although migration should become more effective as the digital preservation community gains practical experience and learns how to select appropriate and effective methods, migration remains largely experimental and provides fertile ground for research and development efforts."[99]

Despite the *Waters/Garrett* note of caution, the RM largely rejects the concurrently proposed solution of *emulation*, and has accepted migration as its solution to the concerns surrounding future information loss. This decision is not without some controversy. Jeff Rothenberg, the most prominent theorist exploring emulation issues, notes that:

> Whereas emulation offers a potential solution to this problem [of the long-term preservation of digital materials], the OAIS dismisses emulation as 'a major technical and economic risk' and assumes (without the support of any empirical evidence) that migration is the most logical preservation approach, despite recognizing that 'digital migrations are time consuming, costly and expose [an archival information system] to greatly increased probabilities of information loss.'[100]

This debate will continue in the digital preservation community. The RM largely supports the concept of migration, so this study operates under that assumption as well. OAIS allowances for data conversion of any sort is a sign that the designers have considered the implications of data storage beyond the immediate short term, and their support for the concept of migration is one of the most preservation-like functions found in the OAIS (even while Rothenberg argues that migration and transformation are not preservation). A significant step for OAIS system design is the development of built-in processes that enable migrations and transformations to be done with some degree of automation.

It is also important to note that the OAIS entities *Administration* and *Preservation Planning* assume completely different roles than the *Management* entity that was earlier modeled as falling outside of an OAIS (along with *Producers* and *Consumers*). *Administration* and *Preservation Planning* both have management responsibilities over various aspects of an OAIS, but they should not be confused with the *Management* entity. In a physical archive situated in a university community, for example, *Administration* and *Preservation Planning* might be represented by a single entity (the University Archivist), while an example of *Management* might be the administrators in the office of the President of the University who have some oversight over the actions of the University Archivist, but only in the most general manner.

5.1.6: Access

Access is the last of the six functional entities covered by the OAIS *Information Model*, and it provides a single user interface to the holdings of the archive. The *Access* entity in an OAIS has no particular dissemination or publication scheme in mind when it makes a digital object available, though particular publication destinations (such as web sites or advertising brochure) are amongst the many possible uses for disseminated objects. The *Access* entity defines three categories of requests that consumers might make of the archive: query requests (through a search utility like OAIster[101]); report requests; and orders, which is a request for delivery of some piece of archival content.

The *Access* function creates a DIP out of an AIP, with dissemination-specific metadata attached. The content, structure and context of the DIP is likely to be different than that of the AIP (the best example being a thumbnail image instead of the archival master) due to the intellectual property or business rules conditions that are likely to be placed on disseminated digital objects.

Access is a familiar term to brick-and-mortar archivists, and the term's usage is consistent in the RM. It is important to note that the *Access* entity is charged with responsibility for the authorization of users hoping to receive requested digital objects. The mapping of functions from the *Access* entity is less important for our current consideration of preservation systems because some manner of access will be built upon almost all OAIS archives. It is important to note the levels of security provided by a particular OAIS implementation, because access restrictions can aid in prolonging the life of archival materials by preventing misuse and the corruption of their content or context. And while an archive would not solely consider current access needs when determining the granularity with which they would store objects in an OAIS, access issues must be considered from the first stages of a design plan. If anything, the designers of OAIS systems must try to predict methods of storage and management of the digital objects over time that will allow the material to be accessible in the distant future. As such, current access conditions may have little impact on the choices that OAIS system designers make. Their majority of design decisions should be made in the *Ingest* entity to gather as much granular information as possible to allow for maximum reuse possibilities down the road.

5.2: The OAIS Information Model

See Appendix B for a detailed view of the Information Model.

As mentioned above, the RM includes both horizontal and vertical components. The horizontal component, the *Functional Model*, represents the workflow in the system, and the movement of archival objects from ingestion through dissemination. The vertical component, the *Information Model*, looks more closely at the elements and layers that comprise each individual digital object, whether represented as a SIP, an AIP or a DIP, and broadly describes the metadata requirements associated with retaining the digital object over the long-term.[102] This exploration of metadata requirements for digital objects in an OAIS explores some of the concepts uncovered when discussing the essential components of the record. The AIP is the item in the repository that features the most complete set of associated metadata, so it is used as an exemplar that also represents SIPs and DIPs.

There really is no single metadata schema that can incorporate the entirety of information that is desired to be captured about items in a repository, but the *Information Model* provides a high-level view of the different types of metadata that needs to be captured for objects in digital repositories. Outside of the RM, metadata experts refer to descriptive, structural and administrative metadata as the complete set of information necessary to fully describe an object. Researchers in the digital preservation community have begun to explore the concept of *preservation metadata*, though this type of metadata may simply comprise a subset of the other three, with preservation as an explicit determination.

The OCLC/RLG Working Group on Preservation Metadata mapped a preservation metadata framework on to the RM, composing an explicit path by which the RM could be implemented in metadata.[103] However, efforts like the Metadata Encoding and Transmission Standard (METS), [104] is a solution finding wider adaptation. METS metadata is expressed using the XML schema language of the World Wide Web Consortium (more on XML in the *Preservation Planning* section). METS is explicitly designed with the RM in mind, and it also attempts to provide a structure by which the complete range of metadata can be encoded in one form. Yet at the same time, METS provides for the development of extension schemas that would allow metadata schemas to be developed for specific resource types (audio, video, etc.). This gives METS both a conciseness and a flexibility that other kitchen sink metadata schemas might not have.

The digital objects preserved within an OAIS are complex, and are composed of layers of information, beginning with the stream of bits at the core of any digital item. However, that stream of bits cannot fully represent all the information necessary to make the object understandable, either now or in the future, and that's why additional layers of information must be stored along with the core bitstream in an archival repository. Each layer of wrapping is made up of descriptive, administrative and/or structural metadata, and the detailed plan for how this metadata surrounding each digital object is organized in an OAIS is the *Information Model*.

The degree to which the Information Model splits each digital object may seem needlessly academic, but as the Information Model is specifically designed to support an implementation strategy (where the Functional Model is clearly conceptual), the completeness of the Information Model takes on useful value.

Each AIP consists of four valuable components. The *Content Information* and the *Preservation Description Information* are the two key components, containing the metadata structures of central importance to preservation functionalities. The *Descriptive Information*, which deals with aids to access, and the *Packaging Information*, which deals with instantiations of archival materials on particular medias, are of less importance to our understanding of a system's preservation needs. These four components and their subcomponents that will be iterated below.

5.2.1: Content Information

The first of the four components of the AIP is the *Content Information.* The *Content Information* represents the core material considered the target of preservation by the OAIS. The *Content Information* comprises two equal parts: the *Data Object*, which is the actual sequence of bits; and the *Representation Information*, which is a combination of *Structure* and *Semantic Information* about the sequence of bits that allows it to be made understandable.

Structure Information helps make the bits understandable "by describing the format, or data structure concepts, which are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc."[105] This *Structure Information* is rarely enough to allow a sequence of bits to be fully understood, however, and that's what makes *Semantic Information* necessary.

Semantic Information represents information such as the language in which the *Data Object* is written, or a simple description of the purpose of the *Data Object*. Scientific data in tables of numbers, for example, would be largely incomprehensible without added *Semantic Information* that explained what the numbers represented.

Semantic Information includes information on the type of operations that might be performed on each data type or the interrelationships between different *Structure* components of the same *Data Object*. The *Structure/Semantic* dichotomy returns us to Cox' definition of the record; a complex object containing content (information), structure (form) and context. Both components are required to make sense of any digital *Data Object*.

Representation Information can also contain references to other *Representation Information* in layers of connection that form a *Representation Network*. In principle, this recursion of connected meaning should continue until a physical representation of the information is found (such as a manual in printed form, or a perhaps a physical copy of the ISO Unicode 4.0.0 standard,[106] which describes the bits that form each of the characters in the Unicode set) at the end of the line, grounding often invisible digital materials in some physical manifestation.

Wrapping the *Data Objects* in layers of *Representation Information* creates complex digital objects encapsulating information on their structure, along with information that describes that structure. This creates objects that to a certain extent can be infrastructure independent, and that are capable of being self-validating and self-instantiating. Needless to say, this makes the task of preserving the *Representation Information* equally important to the task of preserving the *Data Object* itself.

5.2.2: Preservation Description Information (PDI)

The second of the four components of the AIP is the *Preservation Description Information (PDI),* which is focused on describing the past and present states of the *Content Information*, ensuring that is uniquely identifiable and that it has not been unknowingly altered.[107] The *PDI* is describing the types of information that must be preserved if we are to assume that the AIP is preserved, and it is divided into four components: *Fixity Information*, *Reference Information*, *Provenance Information* and *Context Information*. These four components should look quite familiar.

The concepts behind the *PDI* map directly from our earlier discussion of the archival *record*. In fact, the *PDI* section of the RM is taken directly from the sections in the *Waters/Garrett report* that earlier provided support for our definition of record components. Our analysis of fixity, reference, provenance and context in terms of the components of the record apply equally here (the record component of *content* was discussed above in the section on *Content Information*).

Some additional information is helpful to understand how some of these concepts are instantiated in digital repository system design.[108] *Reference Information* identifies mechanisms used to provide assigned identifiers for *Content Information*. In a digital repository, these could include identifiers such as URIs and URNs,[109] Handle System reference assignments,[110] Digital Object Identifiers,[111] or any number of other identification mechanisms. Consistent users of the World Wide Web are aware that some information resources often just disappear. If digital items are to be preserved for the long term, they have to be discoverable. A persistent location identifier helps them to remain that way.

Context Information documents the relationships of the *Content Information* to its environment, including why the *Content Information* was created, and how it relates to objects that exist elsewhere.

Provenance Information documents the history of the *Content Information*, determining any changes that may have taken place since it was originated, and tracking item custody.

Fixity Information provides the data integrity checks or validation and verification to ensure that the *Content Information* has not been altered in any undocumented manner. *Fixity Information* includes special encoding or error detection schemes that are specific to individual *Content Information* objects.

5.2.3: Packaging Information

The *Packaging Information* binds or relates the components of the AIP into an identifiable entity on specific media. For example, if the *Content Information* and the *PDI* are both identified as coming from the same compact disc, the *Packaging Information* would include information about the data structures that comprised the compact disc itself, because this would help to represent how the materials originally were appeared. The *Packaging Information* is not necessarily preserved at any point by an OAIS, but it is definitely not preserved during migration.

5.2.4: Descriptive Information

The *Descriptive Information* is generally derived from the *Content Information* and the *PDI*, and serves as the input to the archival digital objects in the repository by facilitating the development of various access aids, which allow *Consumers* to locate information of potential interest, analyze that information, and order desired information.[112]

6: Concurrent Repository Design Development

The OAIS model helped to explicitly represent archival thinking to the designers of digital repositories for long-term preservation. Thanks to the influence of the OAIS model, much of the present research into digital repository systems, including research coming out of the computer science community, is somewhat informed by archival principles. The completion of the process by which the RM has became an international standard assures that the RM will continue to assert influence on future designers of preservation systems.

However, research and development in digital repositories was taking place concurrently in widely divergent communities of knowledge outside OAIS development. The new language of the OAIS model has proven to be centrally important in enabling divergent knowledge communities to find common ground while working independently on repository designs, but often those designs were developing independently with or without the OAIS model. The influence on the OAIS that developed out of early research done by ARPA, Xerox PARC and other institutions is outside the scope of this paper.[113] It is important to note that many of these efforts, whether they were consciously aware of it or not, were incorporating archival thinking in their design decisions.

One important line of development at technical end of the spectrum began with work from Robert Kahn of the Corporation for National Research Initiatives (CNRI) and Robert Wilensky of the University of California at Berkeley. Their article, "A Framework for Distributed Digital Object Services,"[114] introduced terminology and concepts that would become commonplace in the vernacular of digital libraries and repositories: digital objects; the handle system; and the repository access protocol (that evolved into the OAI-PMH), amongst other concepts.

Their work was picked up by William Arms, also of CNRI, who published his article, "Key Concepts in the Architecture of the Digital Library,"[115] at approximately the same time. Arms suggested the idea that the architecture of the repository should be separate from the content; that the access object is different from the stored object; and the idea that digital objects are more than just the collection of bits.

At the same time, Carl Lagoze of Cornell was incorporating the ideas of Kahn and Wilensky and working on his own secure repository design for digital libraries. His work in computer science eventually evolved into the Fedora architecture, now under continued development at the University of Virginia. Lagoze later helped develop the core services in the Architecture of the National Digital Library for Science Education (NSDL). This NSDL list provides a different perspective on the functional requirements of repository architectures.

These developments in the computer science community spread widely through the library, archives and information science communities through the interest being generated by digital libraries. The developments in repository architectures running through Cornell were merely some of the most prominent in the U.S.: related systems were being developed in a number of countries and in a number of other academic institutions.

As the turn of the century arrived, digital repository system design began to move into the mainstream of research development. The success of the World Wide Web led to a significant redirecting of resources into digital libraries, led by research growth in government and academia. Government research increasingly was led by the LC and NARA and focused on more action-oriented system design. The academic community became energized when the concept of the institutional repository was developed in 2002.

Concurrent with academic and government research specifically directed towards digital repository design, there was a host of research taking place around digital library development (though often completely independent of it) that had a great effect on the development of repository architectures. These include the open source software movement; the development of open standards in file formats; the continued development of metadata standards, networking tools and much more. An exploration of these exterior developments is outside the scope of this paper, but information pertaining to these subjects will be introduced when appropriate.

6.1: Government Efforts

The most well-defined governmental efforts came from LC and NARA, both of whom were charged by law with preserving the recorded output of the United States, and thus had to be proactive about building systems to support that mission. The first tentative steps in requirements documentation were mostly theoretical. These explorations of repository design grew out of the same self-reflective impulses guiding the archival community, and led to a reevaluation of the governmental stewardship mission in the light of changes in digital technology.

Both LC and NARA made efforts to take stock of their digital efforts prior to moving ahead with plans for repository architectures. Commissioned by the Librarian of Congress in 1998, and published in July 2000, *LC 21: A Digital Strategy for the Library* *of Congress*,[116] the on-site study of the Library's technology practices and initiatives, was conducted by a committee of the Computer Science and Telecommunications Board of the National Research Council, and noted the historical developments at the LC while making recommendations for the LC's future technology strategies.[117] In December of 2000, Congress appropriated $100 million for this effort, which instructed LC to develop and execute an approved strategic plan for a National Digital Information Infrastructure and Preservation Program (NDIIP).[118] LC convened a number of national meetings of interested parties, and in October of 2002, published their initial report, *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*.[119] The program plan compiled introductory essays on the various aspects of the digital preservation challenge from researchers operating in variegated domains of knowledge. One essay, "Preliminary Architecture Proposal for Long-Term Digital Preservation" by Clay Shirky[120] described the conceptual framework for supporting the technical functions of NDIIP as a whole, but also touched on some basic functional requirements for a repository. Some of the NDIIP's findings in this regard have since been reexamined,[121] but a pair of items are worth noting. Shirky identified the establishment of unique identifiers for each resource as the principal function of the repository (a requirement that came out of Kahn/Wilensky), and further noted that the repository, as the holder of canonical versions of digital content, should have a focus which included robust security resources.[122]

As the LC was considering a national strategy for digital preservation, they were concurrently working on their own repository design, the Digital Audio-Visual Repository System (DAVRS). DAVRS is the central component of the LC's new multimedia archive being built in Culpepper, Va. We will look at DAVRS more closely in the Case Study section.

Similar efforts were taking place concurrently at NARA (though the funding levels weren't quite as high[123]), represented by the publication of their retrospective *Thirty Years of Electronic Records* in 2003, and their own National Research Council review, *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development*[124] that came out the same year. NARA's Electronic Records Archive (ERA) is a research project in development with the San Diego Supercomputer Center (SDSCC), and will be examined in the Case Study section.[125]

6.2: Academia and Institutional Repositories

While the government was working on both national priorities and individual repository systems, there were a number of activities in academia that began to coalesce around the concept of institutional repositories. In early 2000, the Digital Library Federation (DLF), the Council on Library and Information Resources (CLIR) and the Coalition for Networked Information (CNI) began to address digital preservation questions with a view to facilitating some practical experimentation in digital archiving.

One experiment, funded by the Andrew W. Mellon Foundation, explored the issues surrounding the archiving of electronic journals (e-journals), including the development of archival repositories for e-journals. The issues surrounding e-journals were becoming important in academia due to the growing demand for electronic content, and the concurrently rapid increase in their cost. The e-journal environment also changed the dynamic of control from the institution back to the publisher. Academic institutions discovered that they didn't actually own copies of the journals, but were only being supplied network access to them. The back issues of the journals could disappear if the institution decided to stop paying for current editions, or if the publisher went out of business. The original copies were held on the publisher's servers, and despite protests to the contrary, there was no guarantee that the journals would be preserved for the long-term in the event that something happened to the publisher, or if economic conditions changed for any reason.

These projects, which included work at Harvard, Stanford, Penn, Cornell, Yale, Stanford, M.I.T. and the New York Public Library, explored a variety of issues surrounding electronic journals, with several investigating repository architectures strongly influenced by the OAIS.

One outcome of the early development meetings was a set of minimum criteria for an archival repository of digital scholarly works.[126] This report was strongly influenced by the OAIS RM, and presented a set of criteria for repository design accompanied by a set of directed research issues developed out of those criteria. The central component of their efforts was to focus on the concept of *trust* in digital repository systems. Other important criteria included: gaining physical and intellectual control over the materials to ensure their long-term preservation; developing migration and data validation strategies, along with a scaleable infrastructure; and the development of standards, including standards for preservation and other metadata, standard migration strategies and implementation procedures, standard specifications for physical media and standard accreditation of requirement-conformant archives.

At approximately the same time, the Research Libraries Group (RLG) and the Online Computer Library Center (OCLC) were working on their list of attributes and responsibilities for a trusted digital repository, a set of criteria strongly influence by the OAIS RM, with OAIS compliance was the first criteria by which a repository was to become *trusted*. Trust in digital repositories arises directly out of the archival principle of responsible custody. RLG and OCLC noted that the attributes of a trusted digital repository were very similar to those of an archival entity responsible for reliable custody. In their view, a trusted repositories should include attributes supporting system security, technological and procedural suitability and procedural accountability in addition to the custodial responsibilities. The report then codified these high-level attributes into a set of operational responsibilities that strongly echoed those of the OAIS RM, and which came even closer to a workable list of functional requirements.

As academic research groups explored the ramifications of repository design efforts, momentum was building within the academic institutions to leverage the power of digital repositories towards the preservation of valuable institutional resources. These institutional resources, including web sites, multimedia presentations, electronic pre-prints of scholarly articles and more, had not been previously been considered for preservation, but were now being acknowledgement both for their reuse value, and for their importance as part of the institutional record. The momentum directed towards the management of these resources coalesced around the concept of the *institutional repository* in 2002.

Clifford Lynch, the director of the Coalition for Networked Information (CNI) since July 1997, and a constant presence in digital preservation circles, clearly defined the purposes of institutional repositories in early 2003,[127] and noted that one of the most substantial benefits of the institutional repository movement was that it gave academic institutions a vested interest in the development of digital repository systems. Rather than their previously passive role of supporting publisher initiatives, the institutions were now in the position of supporting research for which they had direct benefit. This self-interest drove repository research and development even harder, leading to the development of several functioning architectures that will be explored in the Case Study section.

7: Commercial Content Management Systems: An Aside

While the research community is addressing the RM and mining it for repository design strategies, an active commercial software market has developed for enterprise-wide systems that address digital authoring, management and publishing problems. These products fall under the general rubric of *Content Management Systems* (CMS), and have been touted to archivists, librarians, and other information professionals as capable solutions to the information collection, management and publishing processes that are too complex to handle informally.[128]

Bob Boiko, a lecturer at the University of Washington Information School, has defined an idealized CMS in terms of its functions,[129] graphing a high-level map of CMS requirements that address some of those of the RM, though the RM terminology and the bundles of functions are slightly different. Boiko's definition divides the functions of CMS products into three main areas: the authoring or acquiring of content; the management of content; and the publishing of content. Additionally, CMS are designed to allow for the flexible reuse and easy manipulation of the materials they contain.

An initial survey of a broad range of CMS product literature has led us to determine that most CMS provide functionalities that comprise an overlapping subset of these three high-level components. It is uncertain to what extent commercial developers have been influenced by the OAIS reference model, but the similarities between the functionalities identified in the OAIS reference model and in the CMS architectures defined by Boiko suggests that the two models are intimately related.

Indeed, the collection, management and distribution of digital resources is something that archivists are interested in, making these products sound very attractive to institutions looking for solutions to their digital management issues. However, market imperatives and business needs have pushed the development of commercial CMS chiefly into the Web Content Management (WCM) area, a distinctly constrained subset of the total range of possible CMS functions, while vendor literature obfuscates the actual functions of the different CMS packages,[130] and systems designed for enterprise-wide installation can entail significant costs.[131] The reality seems to be that each individual product tends to specialize in a subset of the authoring/management/publishing functions, but no product to this point successfully addresses them all in a single product, or within the pricing ranges of most institutions.

Additionally, and most significantly for the digital preservation community, it's uncertain whether any of these products address preservation issues. Mackenzie Smith has suggested that none of them do, staking the development of digital repository systems featuring long-term digital preservation functionalities completely to the research community.[132] There is enough basic conceptual similarity between CMS and OAIS-informed digital repositories to allow for comparison, but we are fairly certain that institutions interested in preservation functionality will have to follow the research path rather than the commercial one at this time.

8: Case Studies

There are but a handful of institutions that have actually implemented digital repositories for long-term preservation. Some larger institutions, such as the LC and ERA, have yet to actively implement a repository, but are significantly involved in the planning for their archives. Other institutions have systems that they've developed from scratch, useful in their own environment, but not extensible outside of it. Other efforts, like MIT's Dspace architecture, are fully designed to be used outside of their home community, and offer great hope for the future in terms of system design.

These projects supply implementation examples that can be utilized to help generate a list of core archival functions that could be found in preservation-oriented information systems. The identification of missing or implemented preservation functionalities could have a number of useful purposes. A list of implemented functions could provide guidance to preservation managers comparing commercial CMS products for possible institutional purchase. It could also assist researchers in identifying areas of strength or weakness in current projects and systems and to guide future research into these systems.

8.1: Harvard Digital Repository Service (DRS) [133]

The Harvard University Digital Repository Service (DRS) developed on an ad hoc basis out of specific institutional needs, and was never designed as a sharable architecture. The design of the Harvard repository system, however, has been very influential on subsequent system design, especially the MIT Dspace architecture, and a number of research efforts surrounding Harvard repository design have made vital contributions to the development of digital preservation infrastructure elsewhere.

The origins of the current Harvard repository system can be traced to July 1998, and the launch of the Harvard Library Digital Initiative (LDI), a five-year program to develop the University's capacity to manage digital information by creating the technical infrastructure to support the acquisition, organization, delivery, and archiving of digital library materials.[134] Jim Coleman, in the Office for Information Systems/HUL, introduced documents that proposed conceptual models for the repository architecture. Many of the elements in these conceptual models would subsequently appear in what became the DRS, and then again in the Dspace architecture.

Coleman's 1998 report, *Towards an LDI Digital Repository*,[135] discussed a number of potential repository services, but maintaining the persistence of the material was always an essential concern. In the process of defining potential services for the repository, Coleman highlighted version control, the automated migration of objects and the management of user and application-specific metadata within the repository as three core services. These services were iterated in more concise form in a later report from March 1999,[136] which laid out the functional requirements of the future repository. These requirements underwent considerable reevaluation during the course of the next few years, but they do provide insight into early design thinking.

"Our repository, in some ways, is a lot closer to Kahn/Wilensky that it is to OAIS," says Dale Flecker, the Associate Director of the Harvard University Library for

Planning and Systems.[137] "That was the dominant paper at the point where we started thinking about this." By May of 2001, the system had a trio of test projects and was ready to accept new submissions from the University community.

The DRS developed out of existing technical infrastructure and was never designed to be a system that could be packaged and shared with other institutions. The DRS is built on an Oracle database structure; Harvard already had a site license with Oracle, and they were less concerned with developing an open source solution than they were in building something with a robust functionality. Additionally, the code for other aspects of the system was written by programmers at Harvard, and has not been made publicly available in the manner of an open source project. At the same time, many of the functionalities that Harvard designed into their system have proven to be quite applicable to other systems, and the range of affiliated research surrounding the DRS development has been quite strong.

Much of this research came about due to the participation of the university in the Mellon Foundation e-journal archiving projects, which commenced in early 2000. These projects were designed to test various aspects of the archiving and preservation of e-journals. The Harvard project demonstrated some concrete methods by which e-journals could be archived, and largely focused on the development of a repository architecture to support the archiving. A significant aspect of this was a detailed model for the technical infrastructure of the repository.

The Mellon grants were explicitly designed to test the OAIS model as a high-level design for a digital repository architecture, and Harvard painstakingly mapped their existing infrastructure to the OAIS model.[138] They also utilized OAIS terminology in describing their repository in their *Report on the Planning Year Grant for the Design of an E-journal Archive*.[139] The *Report on the Planning Year Grant* explicitly stated that the DRS would be responsible for the managed preservation of the deposited objects, but the descriptive metadata associated with these objects was to be stored completely separately in Harvard's extant OPAC system, Hollis.

However, several aspects of their technical design, though not necessarily implemented at Harvard, should be highlighted in terms of their subsequent influence on the design of other repository architectures. The DRS architects identified the need for a central format registry to aggregate all essential information about the formats of materials deposited into the repository. The DRS designers were quite strict about the types of formats that could receive support by the repository, and they designated a list of normative formats that were acceptable. Items submitted into the repository would be converted to one of these formats during the ingestion process; for example, a non-normative format such as jpeg would be converted to the normative format of a tiff. The format registry would include information on the formal name of the format, its version history, a pointer to an authoritative specification, the name of the maintenance organization, the MIME type, technical metadata schemas, compliant tools and validation and migration processes.[140] Their report emphasized a need for a common infrastructure to maintain this type of information. A peripheral project that has developed out of this is the global digital format registry.

The DRS design also implemented the concept of service levels of preservation support. In this concept, different types of materials are guaranteed different levels of preservation management, dependent upon the type and importance of each file format.

This concept was incorporated wholesale into the Dspace architecture design, and it is enumerated in much more detail in the discussion of Dspace. It is important to note that the DRS perceived the highest level of support to include the commitment to monitor formats and associated technologies, to develop and execute migration strategies that attempt to preserve all of the format's native functions and semantic integrity, and to disseminate files in formats that can be rendered by contemporary applications.[141] The format registry and levels of preservation map to traditional archival principles and represent how automated appraisal techniques and the application of levels of control can be applied to digital materials.

"A lot of our work these days is in terms of automated generation of metadata and format validation" continues Flecker. "Jhove[142] is a framework for format characterization and validation. You feed it a file, and if you don't know it's format, it will analyze it for you. It will subtype it within the format, and elucidate its core properties. That's really where a lot of our preservation effort has gone in the last year or so. Jhove is a framework for doing this, and now we're populating it with format-specific information."

The DRS designers also identified the need for persistent identifiers, though their plan was to use identifiers generated locally by their own Name Resolution Service. They noted the problematic nature of using DOIs as persistent identifiers, in that the DOI would resolve to the publisher's resource, as opposed to the unique resource stored in a completely different repository in a likely different form. The DRS designers made plans to include pointers to other identifiers, such as DOIs, in addition to the persistent identifier that they generated themselves.

The DRS designers also identified the need to automate the ingestion process to the greatest degree possible, and they identified METS metadata as the desired "structural envelope" for their SIPs. The DRS researchers focused some of their best research on the structure of the SIPs. Their *Submission Information Package (SIP) Specification*, published in December, 2001, was designed as the authoritative specification of the format of the SIP used by publishers for submission of e-journal content to the archive.[143] and as such, it provided a detailed roadmap, including the types of acceptable formats, and the explicit structure of METS documents, which would be necessary for archival submissions to the DRS. This was the first widely available SIP specification, and still the most detailed, and it supplied explicit guidance for systems designers.

The research related to the development of the SIP specifications also led Harvard researchers to the consideration of Global Format Registry, another piece of essential digital preservation infrastructure. Stephen Abrams, one of the co-authors of the SIP specification, is the driving force behind the Global Format Registry, which seeks to develop an appropriate trust mechanism to encourage the deposit of detailed representation information about proprietary formats, and to make that information available in standard human and machine-readable forms.[144]

Additionally, during the course of the Mellon grant, Harvard commissioned a study by Inera Incorporated to see if it was possible to design a common archival e-journal Document Type Definition (DTD), a document into which Standard Generalized Markup Language (SGML) from a wide range of journal publishers can be transformed such that the resulting SGML preserves and can reasonably render the intellectual content of journal articles.[145] DTDs of this type could be useful in standardizing the structure of archived e-journal materials, aiding the development of common repository architectures into which all of these materials could be deposited. The commission of a feasibility study for e-journals encouraged other knowledge domains to attempt to develop DTDs for their materials as well.

The Mellon Foundation ultimately chose to focus its attentions on the continued development of the digital repository at JSTOR,[146] leaving much of the repository development work at Harvard uncompleted. However, the Harvard researchers made significant inroads into the development of digital preservation infrastructure, and their work on functional system design strongly influenced the development of the Dspace architecture at MIT.

8.2: M.I.T./Dspace[147]

DSpace is an electronic resource management enterprise at the Massachusetts Institute of Technology (MIT), closely related to MIT's Open Courseware (OCW), an electronic publishing initiative designed to ”provide free, searchable, coherent access to MIT's course materials."[148] Both DSpace and OCW (along with OKI, the Open Knowledge Initiative) are part of the large infrastructure initiatives organized by the MIT Council on Educational Technology (MITCET). Initially funded in 2000 as a program of the Invent@MIT joint venture created by Hewlett-Packard (HP) and MIT,[149] the development of the Dspace architecture is now overseen by the MIT libraries. There is a strong connection between the work done by Harvard researchers on their DRS and the Dspace project at MIT. Mackenzie Smith, currently the Associate Director for Technology, MIT Libraries, and a member of the Dspace development team, was the

Digital Library Projects Manager at Harvard between January 2001 and December 2001, the period when Harvard was implementing their Mellon e-journal grants.

Within the broad mandates provided by MITCET[150] DSpace has their own mission: to provide stable long-term storage to house the digital products of MIT faculty and researchers; to provide long-term preservation for digital materials in a variety of formats, including text, audio, video, images, datasets and more; and to enable remote access to those materials through one coherent interface.[151] Long-term preservation functionalities have been an explicit attribute of the system from its initial stages of design, and Dspace is the current de facto repository architecture standard. This is due to a variety of social and political reasons in addition to its advanced design.

Unlike the Harvard DRS, Dspace was first to market with a complete end-to-end system, and they were able to successfully leverage the power of the open source software development community to build early interest in the system as a freely available tool. MIT had solid funding from HP for development, and was also able to leverage their name recognition to generate substantial interest in the Dspace architecture while other systems were still in the pre-operative phase. These social advances have created significant interest in the Dspace architecture, and an active community of developers, implementers, researchers and users has grown over the past two years which might possibly sustain the development beyond the initial interest. Other experimental repository architectures have not been able to gain this level of uptake. This implies that researchers will first consider the Dspace architecture when making their own build-or-buy decisions, and promises to keep Dspace in the mix for the foreseeable future.

In addition to its primary use as a repository for MIT faculty materials, DSpace is designed to facilitate research in digital content-management systems and the related issues of preservation, archiving and distribution of digital materials. The DSpace platform is explicitly informed by the Open Archival Information System (OAIS) reference model. As has been demonstrated, the OAIS reference model provides a thorough vocabulary for describing media archive systems, and for crosschecking the functional and operation plans for a proposed archive.[152] This direct OAIS-influence on the design of the Dspace architecture allows it to provide the first and best example of archival-influenced functionalities in repository design. This makes the Dspace architecture an appropriate baseline by which other architectures can be compared and contrasted. An overview of the most significant architectural features of the Dspace design provides an appropriate overview to the possibilities of archivally-influenced repository design.

Dspace is envisioned as a general purpose repository, which implies that its functionalities must be able to apply to the complete range of potential digital materials. Here, distinctions are made between bit-level preservation, where the bit order of each item is preserved, and the ability of a system to preserve items at a higher level of representation and understanding. These higher-level abilities touch on the archival representation issues discussed earlier in terms of the sanctity of evidence and the components of a record. The preservation of these levels of understanding is more difficult and challenging for repository systems.

The Dspace system is designed to handle a multitude of file types, and will ultimately incorporate the materials from a large percentage of faculty and researchers at

MIT. An understanding of how the Dspace system promotes preservation assumes an analysis of the materials housed in DSpace (the content) and the accompanying processes that organize, document and disseminate them (the services).These can be divided into *content types* and the *submission process*.

DSpace uses the terminology *communities* to define content spaces within which users can browse for information. Within each of these communities, users can then access *collections* of information. Under collection we find *items* (described in the DSpace documentation as the "archival atom"; "a grouping of content and metadata that ...makes sense to archive as a single unit."[153]). The *item* in Dspace parlance is the equivalent to the definition of *record* found here. Items are organized into *bundles*, a grouping of bitstreams, such as might be found with an html file and its associated image files, and at the lowest level, the items in DSpace are organized by *bitstream*. Each bitstream is then identified by its *bitstream format*:

> This is a set of information...describing as much as we know about the format and encoding of the bitstream, including MIME type and name of the type (e.g. "Adobe PDF"). This may also hold information such as the specification of the format, and source code for manipulating the format.[154]

This hierarchical division of bitstreams and their associated groupings or added metadata roughly corresponds to the OAIS *Information Model*.

The preservation of these different content types is managed in the Dspace architecture by the Bitstream Storage Manager. The Bitstream Storage Manager is one of the most significant preservation functionalities of the Dspace architecture, and stores information about the submissions at the bitstream level and provides a limited transactional capability for the bitstreams. The system prescribes different *levels of support* for each set of associated bitstreams. At the lowest level of support the Storage

Manager will remember the sequences of bitstreams associated with each submitted item, a level of service that is relatively simple to maintain. Complexity is introduced as the Storage Manager and its administrators attempts to identify a preservation service level for each of the items, and then determine the degree of support provided for each of those bitstream types.

DSpace administrators divide the bitstream formats into two categories: *Known* and *Unknown*. *Known* bitstream types are further subdivided into formats that are either *Supported* or *Unsupported*. *Known* means that:

> DSpace administrators have named the file format, its version, mimetype, and any other relevant information (for example, relationships with other formats) in the system's bitstream format registry; and a process is in place to actively identify incoming submissions using the named format.[155]

Supported means that:

> DSpace administrators have procured and stored sufficient specifications documenting the format – either within DSpace or in analog format within a trusted library.[156]

The administrators knowledge of, and support for, various bitstream formats culminates in different levels of preservation service. The top level is referred to as *Level 2, Supported*, and is characterized by the following:

> The bitstream will be maintained and returned upon request in the future. Additionally, the host institution believes that the ability to use and understand the submitted material can be reasonably preserved into the future, and commits to exercise its best effort to do so, in a manner appropriate to the material's context. Preservation techniques may include emulation, migration, transformation services, or other strategies.[157]

The functionalities of the Bitstream Storage Manager fall most clearly under the OAIS conception of Preservation Planning, but there are also clearly elements of

Administration and Archival Storage at play here as well, which highlights how broadly the OAIS model can be applied across functions. The identification and management of preservation support levels for digital information is one of the key components of the Dspace architecture that the MIT designers borrowed from the Harvard DRS repository design. Support for present and future file types, and planning for the adaptation of those same files, is one of the key issues facing digital libraries. The DSpace model is an example of how a trusted repository can ensure the maintenance of widely disparate file types by planning in advance for future maintenance, and identifying, early in the preservation process, formats and files that it is capable of supporting. This enables an organization to highlight endangered files earlier in the preservation process, and also provides a coherent plan for organizations to retain copies of software, hardware and documentation that will provide support to identified file types over time. This prescient identification of future file support needs ensures contributors that their work will receive preservation care.

This transparent preservation planning helps to imbue the repository with trust. When contributors receive assurance that, at minimum, the bitstreams of their files can be supported over time, their confidence in allowing the repository to control their materials increases. Contributors are asked to grant a non-exclusive license to DSpace to distribute the contribution and to translate it for the purpose of preservation. Because license terms change over time, DSpace stores a copy of the license as a bitstream within the item so that the specific terms agreed upon are always available.[158]

Trust is also built into digital systems through security measures. These security measures also help to ensure that the materials retain their authenticity and reliability over time. Anyone can browse and search the material in the DSpace repository at MIT, but in order to contribute materials the user must be registered. DSpace refers to registered users as "e-people" to reflect that the users may be machines as opposed to actual people.[159] Verification of registered user status includes the authentication of username and password information for individuals, or an IP-address-based network presence for machines. Digital signatures, in the form of X.509 certificates, also have support. In order for a user to perform an action on an object, they must have permission. There are several different levels of permissions (or "actions") in DSpace: Read; Write; Add; Remove; and Workflow:

> **Read**: The action of knowing of an object's existence, and viewing any metadata associated with it.
> **Write**: Modifying the metadata associated with an object. This does not include the ability to delete.
> **Add**: The action of adding an object (e.g. an item) to a container (e.g. a collection). In order to submit an item to a collection, an end-user must have ADD permission on that collection.
> **Remove**: The action of removing an object from a container.
> **Workflow**: May participate in a workflow associated with a collection; for example, permission to reject a particular submission from entering the collection.[160]

Submission in DSpace by users with the proper permissions operates at three levels: content, metadata, and bitstream, though these levels are not necessarily discrete. In line with the OAIS model, submission to a Dspace repository is referred to as the *ingest process*. The *ingest process* can begin in one of two ways. Multiple items can be introduced by a batch-item importer, which turns an external *SIP* (an XML metadata document with some content files) into an *in-progress submission object*. Or individual users can enter data through the web submission user interface to create the same type of in-progress object. Users create both the intellectual content and the metadata for each submission they create. Users specify the baseline metadata for each item they submit through a common end-user form that is required for all submissions.

Baseline metadata for individual items in DSpace is based on the fifteen descriptive fields in the Dublin Core (DC) metadata schema. More specifically, DSpace utilizes an adapted version of the Dublin Core Library Application Profile (DC-Lib), one of a series of DC schemas optimized for a particular local application.[161] The DC metadata schema is most useful for providing descriptive metadata for each of the objects to which it refers, but its relative simplicity allows it to easily map to other schemas, making it highly adaptable for an uncertain future. The Dspace designers are also exploring METS metadata as the next step in providing more complete metadata coverage, and an affiliated MIT research project, SIMILE, is ranging even further afield to determine ways to combine numerous metadata schemas into one descriptive architecture.[162]

When the contributor has completed their work on the content of the item for submission, as well as having completed the addition of metadata elements, the object fully enters the workflow process. While the term "workflow" can apply broadly, DSpace most specifically refers to the editorial review process as "workflow." Submissions in DSpace almost universally go through a human-coordinated editorial review process, which generally takes place as close to the individual communities as possible. In addition to the layers of human-coordinated review, there are a number of automated processes that apply metadata to the submitted item. These processes are overseen by the History System functionality of the Dspace design. The History System functions most explicitly represent the archival principles of provenance in the digital repository. The History System:

> Assigns an accession date;
> Adds a "date.available" value to the Dublin Core metadata record of the item;
> Adds an issue date if none already present;
> Adds a provenance message (including bitstream checksums);
> Assigns a Handle persistent identifier;
> Adds the item to the target collection, and adds appropriate authorization policies;
> Adds the new item to the search and browse indices.[163]

The provenance message includes filenames and checksums of each file that are uploaded with the item, and that can be used by DSpace administrators and users to verify the integrity of the content and metadata within the system.[164] Each time the workflow process changes the item, another provenance statement is added. Changes in the item also invoke the History system, which creates Resource Description Framework (RDF) data describing the current state of the object. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web.[165]

Also of interest is the persistent naming technique that DSpace employs utilizing a Corporation for National Research Initiatives (CNRI) Handle server. The Handle System is one of several persistent naming systems that have achieved substantial institutional uptake. DOIs have been mentioned previously, and these would also include OCLC's Persistent Uniform Resource Locators (PURLs).[166] Efforts are underway by Dspace architects to provide users with the flexibility to incorporate different Uniform Resource Name (URN) systems if desired. Each submitted item in DSpace is given a persistent URN that is managed by the Handle server. This will enable the item to be found over time despite changing physical locations:

> The Handle System...protocols enable a distributed computer system to store handles of digital resources and resolve those handles into the information necessary to locate and access the resources. This associated information can be changed as needed to reflect the current state of the identified resource without changing the handle, thus allowing the name of the item to persist over changes of location and other state information. Each handle may have it own administrator(s), and administration can be done in a distributed environment. The name-to-value bindings may also be secured, allowing handles to be used in trust management applications.[167]

Once these automated and human-editorial processes are complete the submission process archives the item in DSpace as an AIP. The DSpace system utilizes a number of different types of advanced technology to facilitate its processes. Some of the most interesting technologies, including the Handle System, provenance binding with bitstream checksums, and the bitstream storage manager have already been briefly enumerated, but it is important to note that the DSpace system operates entirely within the open-source software framework, freeing it from the encumbrances of proprietary software. Building the system entirely on open source tools is one way that the Dspace designers are attempting to ensure the viability of the architecture over the long-term. While there is no guarantee that open source software products will last any longer in the marketplace than their commercial counterparts, there is a wide body of literature suggesting that the open source software development model can produce materials as robust as commercial products, and that the social network of innovation surrounding open source tools can produce a sustaining community for those tools.[168] Dspace acknowledges the power of the open source model by making its source programming software freely available under the terms of the Open Source Initiative's BSD license.[169] The majority of the prerequisite software for running the DSpace system is also open source, including a UNIX-like operating system, the Apache web server, the Tomcat servlet engine, the PostgreSQL database system and others.

8.3: National Archives and Records Administration's Electronic Records Archives (ERA)[170]/San Diego Supercomputer Center's Storage Resource Broker (SRB)

Like many United States government projects, the development of the design of the NARA Electronic Records Archives (ERA) has been largely transparent, yet incredibly complex. Both NARA and the Library of Congress are meticulous in documenting each step that they take in repository design, and this documentation can be incredibly useful to researchers who follow. In many instances, it is expected that NARA and the LC will take the lead on projects of this sort, though the institutional repository movement is out of the gate with a functioning digital repository design in the Dspace architecture, while both NARA and the LC are still in the development stages. The NARA ERA has not been built, nor is it in the production phase. However, there is a rich body of materials available that document the research actions taken to date, and the development of requirements for the archive that are very useful in understanding how the archival profession translates its needs into the design of a digital repository for long-term preservation.

NARA's historical involvement with electronic records has been complicated by wide fluctuations in funding and rapidly shifting political winds. These historical efforts are detailed in the *Thirty Years of Electronic Records* book, which covers the most significant aspects of NARA's involvement in electronic records, and also provides a brief history of the earliest stages in the development of their electronic records archive, the impetus for which came out of the need to preserve almost 40 million email messages from the Clinton administration. Despite NARA's more than thirty years experience in electronic records management, it was clear that they did not have the functional capabilities or quantitative capacity to handle those types of material at that time.[171] This awareness led directly to the formation of the NARA ERA. The first step in negotiating a development plan was to undertake a survey of other government agencies to determine whether there were other electronic records programs being implemented from which NARA could gain insight. They were eventually led to the high-performance computing infrastructure. One particular project, the Distributed Object Computation Testbed (DOCT), being overseen by the San Diego Supercomputer Center (SDSCC), was exploring an environment for handling complex documents (in this instance, digital patent application case files) on geographically distributed data archives and computing platforms. These patent applications were interesting to NARA in that they were complex digital items that included graphics in addition to the text, and they most clearly replicated the structure of the items found at NARA. The DOCT project was not explicitly researching the long-term preservation aspects of the patent applications, however. The interest in preservation concerns led NARA to propose a direct collaboration with the SDSCC on a testbed of materials designed to explore long-term preservation functionalities.

Both NARA and the SDSCC were concurrently involved in the process of developing the OAIS model. NARA was also involving itself in a number of other partnerships at this time to explore further aspects of electronic records management. These included the PERPOS project, a collaboration between NARA, the Georgia Tech Research Institute and the U.S. Army Research Laboratory to develop automated tools for identifying the data type of each file in a system; and the Interpares project discussed earlier. NARA also became a member of the Digital Library Federation, and worked closely with the LC on the development of their NDIIP program. These collaborative projects laid the groundwork for the construction of the NARA ERA, but there were still a number of difficult issues for NARA to confront.[172]

The ERA would essentially be a leading-edge production digital-archive system, as there were no off-the-shelf systems available to handle the quantity of information that NARA would be dealing with. The National Academies report on the ERA project was somewhat critical of the agency's preparation through early 2003, suggesting that the experiments conducted with the SDSCC did not fully explore questions regarding the scalability, complexity, trustworthiness or operational details of a production repository system. Additionally, the report questioned the strategy of migrating the records to an XML-based format, asserting that these attempts to circumvent the semantic constraints within records were still in the experimental stages and far from ready for inclusion in production systems,[173] a strategy that had been advanced in a number of other research programs in order to obtain the format-independence of semantic material.

NARA has recognized that it doesn't have the in-house capabilities to design the ERA itself, so it prepared a Request for Proposal (RFP) in December of 2003 designed to state the high-level requirements for that ERA that would be further decomposed into a set of systems requirements in collaboration with the contractor. As of April 2004 the bidders list included Accenture, CSC, Harris, IBM, Lockheed Martin, Northrup Grumman, Optimal Solutions & Technologies, SAIC, and Vecna Technologies

Inc.[174] The RFP explicitly states that the system must support archival processes for such activities as appraisal, scheduling, and description that apply to both electronic and non-electronic records, as well as capabilities for the automated archival processing of electronic records themselves, and it uses the structure of the OAIS RM to frame their discussion of each functional area. Particular archival processes include:[175]

- Physical transfers of sets of electronic records, via telecommunications and on physical media, for ingest into ERA;
- Verification that transferred sets of electronic records conform to disposition agreements;
- Validation of the representation information for any set of electronic records;
- Long-term storage of electronic records;
- Transformations of electronic records to maintain accessibility and authenticity;
- Characterization of electronic records for archival description;
- Redaction of restricted content;
- Search, retrieval, presentation, and output of the records; and
- Disposal of records authorized for destruction.

Additionally, the system would eliminate or minimize records' dependence on any specific hardware or software while maximizing the types and sources of electronic records and digital data created using any type of application on any computing platform.

NARA additionally expressed some developmental constraints that would apply to the design of the ERA. The concepts behind these constraints were developed out of NARA's earlier research into electronic records, but especially out of their work with the SDSCC:

- ERA will transform electronic record/data types into a hardware and software independent format
- ERA will express representation information in XML format
- ERA will exchange representation information consistent with the Metadata Encoding and Transmission Standard (METS) version 1.3.
- ERA will export self-describing media containing electronic records from the ERA primary data storage repository

- ERA will import self-describing media containing electronic records into the ERA primary data storage repository
- ERA will manage electronic records according to the access restrictions of the record
- ERA will store and process electronic records in environments appropriate to their stated access restrictions
- ERA will prohibit unauthorized alteration of electronic records
- ERA will control access to electronic records in accordance with the records' access restrictions
- ERA will prohibit redaction of the preservation copy of an electronic record

Several of these points ran counter to the suggestions from the National Academies report, suggesting some disagreement between the Academy reviewers and the NARA management on the future directions of the ERA.

Several of the points above deserve a bit of amplification. Some discussion has already been made on the subjects of data security, the use of XML for format independence, and the increasingly widespread adoption of the METS metadata schema. NARA has been involved in the development of each of these functional areas, but their work with the SDSCC on persistent archive research, and in the development of self-validating, self-instantiating knowledge-based archives deserves special attention.

> For permanent records – those preserved forever – as well as for some temporary records which need to be kept for lengths of time that exceed several generations of information technology, it will be necessary to transform the records from the formats in which they were received to persistent formats. A persistent format is one that is supported by a preservation strategy for diminishing the impacts of technological obsolescence, minimizing dependence on specific hardware and software, and enabling retrieval and output of authentic copies of the records in the future. An ideal persistent format would be self-describing and be able to be validated in accordance with open, non-proprietary standards.[176]

The preservation of the context associated with digital objects is the dominant issue for collection-based persistent archives, and the goal is to store the digital objects comprising the collection and the collection context in an archive at the same time. The re-creation of the data is done with a software program that uses pre-defined schema descriptions to generate the collection, with the ultimate goal the development of a generic program that works with any schema description. These persistent archives are inherently composed of heterogeneous resources, and the challenge is to design data handling systems that provide the ability to interconnect archives with databases containing information that describes the archival materials. The data handling system at the SDSCC is called the Storage Resource Broker (SRB). The SRB supports the protocol conversion needed for an application to access data within either a database, a file system or an archive. The information the driver uses to access a particular data set is maintained in the associated Meta-data Catalog (MCAT), a database containing information about each data set that is stored in the data storage system. One of the main purposes of the SRB is to provide uniform access to diverse storage resources in heterogeneous computing environments. To that end, the SDSCC has been experimenting with the development of computer grids; large collections of heterogeneous storage devices that are networked together through the use of software like the SRB, and that can provide unsurpassed storage and computing power.[177] The thinking behind self-instantiating data concept came slightly later:

> To achieve the goal of reinstantiating archived information on a future platform, it is not sufficient to merely copy data at the bit level from obsolete to current media but to create "recoverable" archival representations that are *infrastructure independent* (or *generic*) to the largest extent possible. Indeed the challenge is the forward-migration in time of *information and knowledge about* the archived data, i.e., of the various kinds of metainformation that will allow recreation and interpretation of structure and content of archived data ... *self-validating* means that declarative constraints about the collection are included in *executable form* (as logic rules). [author italics].[178]

These two concepts were significant contributions on the part of the NARA and SDSCC researchers to the theoretical discussion of repository design, though it remains to be seen how they will be instantiated in systems design. These theoretical aspects are certainly represented in the NARA RFP as systems requirements.

The RFP can be considered the most evolved statement on the functional requirements for a digital repository architecture that supports long-term preservation, yet it is still a requirements document and doesn't yet represent design solutions that have been forged in practice. It will be several years before the NARA requirements begin to appear in a functioning system.

8.4: Library of Congress (LC) Digital Audio-Visual Repository System (DAVRS)

The efforts at the LC to design a digital repository system have paralleled those at NARA. A brief history of the LC's involvement in digital preservation issues was iterated earlier, centering around the publication of the report *LC 21: A Digital Strategy for the Library of Congress* in 2000. This report briefly referenced LC's experimentation with the Artesia TEAMS content management system, but noted that the TEAMS system may not contain all the functionality necessary for LC to fulfill its preservation duties. LC also had its own storage issues. In December 1997, the Congress authorized the acquisition of space in Culpepper, Virginia owned by the Federal Reserve Bank of Richmond to be used as the National Audio-Visual Conservation Center (NAVCC). The center was projected to be sufficient to house all of LC’s audio visual collections for the next 25 years. There were deadlines, however: Congress had approved the management development plan for the Center that enabled full occupancy no later than 2005.[179]

The plan to develop the Culpepper site was one impetus for LC research into the development of a digital repository system that would be used for its audio-visual materials. As with the NARA ERA, to this point no function system has been developed by the LC. In the same manner as the ERA, however, a rich body of theoretical materials and requirements documents have been created that show LC moving slowly but surely to some sort of finished product.

In the same way as NARA, the LC's involvement in basic research in a number of areas and at a number of levels has contributed to the development of digital repositories. The requirements definitions that have developed around the design of the Culpepper DAVRS system is one area in which the LC has made very specific design contributions. A second very significant contribution has been their work on the development of the METS metadata schema.

The DAVRS repository system requirements had been elucidated in a Systems Requirements Document, LC-DAVRS-02, from March 2000. This document, along with the Conceptual Design Document, LC-DAVRS-03, were both available as of February 2003, but have now disappeared from the LC's web site. LC-DAVRS-03 provides a rich conceptual design of the future LC repository, and represents an early stage view of what the system designers imagined the DAVRS system might look like, from a number of different angles. LC-DAVRS-02 is essentially a laundry list of requirements that should be found in the system.

The lengthy requirements list in LC-DAVRS-02 is ordered by OAIS functional area, and as such, provides an exceptional resource for the direct mapping of OAIS concepts to specific preservation tasks in a digital repository architecture. Outside of this pair of documents, there is little other publicly available material describing the specifics of the DAVRS repository design. It appears that a majority of the designer's attentions have been directed towards their consideration of the METS metadata schema.

The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the LC,[180] and has been a significant contribution to the development of metadata requirements for archival objects in digital repositories. While the METS initiative did not arise directly out of work with the DAVRS, much current research surrounding the development of the LC repository has centered on developing and refining the METS schema. The details of the METS schema are too complex to explore in any great detail here, but a high-level introduction notes that the schema is divided into seven sections:[181]

- METS Header - The METS Header contains metadata describing the METS document itself, including such information as creator, editor, etc.

- Descriptive Metadata - The descriptive metadata section may point to descriptive metadata external to the METS document (e.g., a MARC record in an OPAC or an EAD finding aid maintained on a WWW server), or contain internally embedded descriptive metadata, or both. Multiple instances of both external and internal descriptive metadata may be included in the descriptive metadata section.

- Administrative Metadata - The administrative metadata section provides information regarding how the files were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of the files comprising the digital library object (i.e., master/derivative file relationships, and migration/transformation information). As with descriptive metadata, administrative metadata may be either external to the METS document, or encoded internally.

- File Section - The file section lists all files containing content which comprise the electronic versions of the digital object. <file> elements may be grouped within <fileGrp> elements, to provide for subdividing the files by object version.

- Structural Map - The structural map is the heart of a METS document. It outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element.

- Structural Links - The Structural Links section of METS allows METS creators to record the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map. This is of particular value in using METS to archive Websites.

- Behavior - A behavior section can be used to associate executable behaviors with content in the METS object. Each behavior within a behavior section has an interface definition element that represents an abstract definition of the set of behaviors represented by a particular behavior section. Each behavior also has a mechanism element which identifies a module of executable code that implements and runs the behaviors defined abstractly by the interface definition.

METS has found adoption because it is able to provide an envelope to contain the variety of rich metadata necessary for the preservation management of digital objects

An LC document related to the development of metadata schemas for the DAVRS project was LC-DAVRS-07 from January 2001, a paper on the metadata requirements and basic structure of the Archival Information Package (AIP) as envisioned by LC repository designers. This document is a detailed analysis of the metadata components that would be attached to an AIP stored in the digital repository. This reports suggests ways that the METS schema, which was being developed concurrently with this project, could be extended to include additional metadata that would serve long-term metadata needs. The current LC work has been focused on developing extensions to the METS metadata schemas that more precisely define the metadata necessary for the audio and video files that the LC is attempting to preserve.[182]

8.5: Other Repository Developments

There are a number of other significant research projects exploring digital repository or digital preservation issues that are peripherally affiliated with the design of repositories for long-term preservation, but time and space constraints prevent a more detailed exploration. These include the Fedora architecture, originally conceived at Cornell and now being developed as a sharable digital repository by the University of Virginia. The research path of the early development of Fedora was iterated earlier.

Fedora uses an object-oriented programming approach to define data objects, and is designed to operate completely within the web environment. The original Fedora research implementation was built in a distributed object paradigm using the Common Object Request Broker Architecture (CORBA). The Virginia reinterpretation is proving that the model can be adapted to run as a web application, specifically using Java Servlet technology with relational database underpinnings. However, the early prototypes sacrificed some of the advanced interoperability features of Fedora. The current version of Fedora recreates a full-featured Fedora system that can become a foundation upon which interoperable web-based digital libraries can be built. However, Fedora is not explicitly confronting digital preservation issues, and is focused more on building digital objects that have intelligence built into them. It has also developed with little archival influence. Other repositories of note include those being developed at WGBH in Boston; the repository development surrounding the California Digital Library; the Texas Email Repository Model (TERM) at the University of Texas; and the JSTOR Electronic-Archiving Initiative.

9: Lessons Learned

This research project has explored the ways in which traditional paper-based archival principles and practices, based on the requirements of a knowledge system built on physical forms (papyrus, parchment, paper), are being applied to the development of repositories designed explicitly for the long-term preservation of digital materials. Though debate remains active, the archival community has gradually coalesced around a set of high-level principles and practices generally agreed as representative of the core values of archival activity: the sanctity of evidence; the preservation imperative; the primacy of the record; respect des fonds, original order and provenance; and hierarchy in records and their collective description.

Disparity can still be found between archival institutions in their implementations of archival principles and practices. The development of the World Wide Web has presented innumerable opportunities to archival institutions to widely distribute information about their archival holdings. This distribution has been facilitated by the development of common information exchange procedures and standards that have forced institutions into common practices. The pressures of conformance have led archivists to reevaluate their principles and practices in elemental ways and decide whether these principles and practices can be adequately applied to the preservation of digital materials. These principles and practices, developed over the course of centuries, do not simply dry up and blow away in the face of the challenge of digital materials. Archivists have traditionally been tasked with defining and promoting the social utility of records and to identify, preserve, and provide access to documentary heritage regardless of format, and this traditional role is equally provident in the consideration of digital materials.

I have been able to show that traditional archival principles are being incorporated into preservation system design through the mediating effects of the Open Archival Information System Reference Model. The OAIS model, developed by the space science community but with active contribution from the archival profession, is a high-level model for information system design based on archival principles. The continued implementation of the OAIS model is a major influence on the acceleration of archival influence on information system design.

The OAIS model's high-level view of archival system design is most successfully applied as a common language bridging the gaps between different knowledge domains. Information system designs are increasingly described utilizing the OAIS language, and the level of understanding achieved between records managers, computer scientists, government agencies and academic researchers has increased concurrently to the benefit of all. The case studies above document this cross-domain involvement, but they concentrate on demonstrating how the OAIS model has increased the archival influence on repository design.

A significant lesson of this research is an acknowledgement that digital repositories for long-term preservation are still in the early stages of development. The development of these types of systems is being driven almost entirely by the research community, as there is currently little preservation functionality represented in the commercial content management business sector. There are no out-of-the-box solutions to digital repository issues at this point, and it may be five years before the most promising research systems are robust enough for their deployment in production environments.

The research presented here has been to demonstrate that archival principles and practices, mediated by the OAIS reference model, are finding their way into the design of currently functioning digital repository architectures. The Harvard DRS designers included functional components of their system designed to support persistence over time. A pair of Harvard DRS concepts developed in the course of the Mellon-funded e-journal grant, (the format registry and levels of preservation support), would find wider application in MIT's Dspace architecture.

The Dspace designers were able to incorporate many of the significant components from the Harvard DRS, then leverage the power of the open source software development model to build support for their system, which was released as an end-to-end product in November 2002. The wide availability of the Dspace software has encouraged a significant number of institutions to get involved in its continuing development, which bodes well for the future viability of the platform. Dspace project development is important on a number of levels. Dspace has compellingly instantiated research concepts into a piece of widely available software, which has then facilitated further discussion throughout the archival and information science community (both pro and con) on implementation decisions made by the Dspace designers. This ongoing discussion should prove to be beneficial to the archival community in developing future robust repository architectures.

The efforts being made at NARA and the LC have developed more slowly than the Dspace architecture, but have incorporated knowledge of the above two projects in their development. Due to the strict production requirements of both NARA and LC, the systems they design may prove to be more precisely developed and durable than the Dspace architecture, though perhaps not as flexible. The government, should it choose to continue to acknowledge the importance of digital archiving, can structure the terms of debate in elemental ways. The rigorous design methodologies undertaken at NARA and the LC have been slow in developing, but they supply a detailed set of tools by which researchers can apply the lessons learned to their own future explorations of archival principles and practices in the design of digital repositories for long-term preservation.

Additionally, there is a huge range of affiliated issues surrounding the development of digital repositories that must be confronted. These issues include the open source software movement; the development of open standards in file formats; the continued development of metadata standards, networking tools and much more. Developments in these affiliated area will have a profound effect on the direction of repository design geared toward the long-term preservation of digital materials.

10: Notes

The Web site addresses listed in this section were valid as of as May 2, 2004.

[1] John Garrett and Donald Waters, eds., "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information," May 1996: 3-4. Available at http://www.rlg.org/ArchTF/.

[2] Anne J. Gilliland-Swetland, "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment," Washington, D.C.: CLIR, 2000. Available at http://www.clir.org/pubs/reports/pub89/pub89.pdf.

[3] An introduction to Schellenberg's work can be found in T.R. Schellenberg , *Modern archives: Principles & Techniques*, Chicago: Society of American Archivists, 2003.

[4] Patricia Kay Galloway, Assistant Professor, Archival Enterprise and Digital Asset Management, School of Information, University of Texas-Austin. Interview by author, 27 February 2004, Austin, Tx. Tape Recording

[5] *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report*. Mountain View, CA: Research Libraries Group. 2002: 3. Available at http://www.rlg.org/longterm/repositories.pdf.

[6] See "World Wide Web Consortium." Available at http://www.w3.org/.

[7] See "SAA History." Available at http://www.archivists.org/history.asp.

[8] "The term 'archive' in the name Open Archives Initiative reflects the origins of the OAI – in the E-Prints community where the term archive is generally accepted as a synonym for repository of scholarly papers. Members of the archiving profession have justifiably noted the strict definition of an 'archive' within their domain; with connotations of preservation of long-term value, statutory authorization and institutional policy. The OAI uses the term 'archive' in a broader sense: as a repository for stored information. Language and terms are never unambiguous and uncontroversial and the OAI respectfully requests the indulgence of the professional archiving community with this broader use of 'archive'." See the "Open Archives Initiative FAQ. Available at http://www.openarchives.org/documents/FAQ.html#What%20do%20you%20mean%20by%20an%20%22 Archive.

[9] For a succinct introduction to archival history see James M. O'Toole, *Understanding Archives and Manuscripts,* Chicago: Society of American Archivists, 1990: 27-47.

[10] Maygene F. Daniels and Timothy Walch, eds., *A Modern Archives Reader : Basic Readings on Archival Theory and Practice*, Washington, D.C. : National Archives and Records Service, U.S. General Services Administration, 1984: 339.

[11] Daniels and Walch, 341.

[12] Fredric M. Miller, *Arranging and Describing Archives and Manuscripts*, Chicago : Society of American Archivists, 1990: 16. I touch only briefly on the myriad ways these materials are different from each other, choosing instead to focus on the ways in which they are similar. Miller builds an even better case for the commonalities between the three types of materials.

[13] James M. O'Toole, *Understanding Archives and Manuscripts,* Chicago: Society of American Archivists, 1990: 42.

[14] Steve Henson, *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries*, Washington D.C.: Library of Congress, 1983: 1.

[15] "The Dublin Core metadata element set is a standard for cross-domain information resource description. Here an information resource is defined to be 'anything that has identity'. This is the definition used in Internet RFC 2396, 'Uniform Resource Identifiers (URI): Generic Syntax', by Tim Berners-Lee et al. There are no fundamental restrictions to the types of resources to which Dublin Core metadata can be assigned." Available at http://www.dublincore.org/documents/dces/.

[16] See "Encoded Archival Description (EAD)" at http://www.loc.gov/ead/.

[17] The elucidation of these principles is strongly influenced by Gilliland-Swetland, who provides the baseline for analysis, but the categories do not conform exactly to hers, and ideas and concepts found elsewhere are also incorporated.

[18] Gilliland-Swetland, p. 10.

[19] Hilary Jenkinson, *A Manual of Archive Administration*, London: Percy Lund, Humphries and Co., 1966: 4. He later acknowledged the primacy of evidential value in his influential "Reflections of an Archivist" article. See Maygene F. Daniels and Timothy Walch, eds., *A Modern Archives Reader : Basic Readings on Archival Theory and Practice*, Washington, D.C. : National Archives and Records Service, U.S. General Services Administration, 1984: 15.

[20] The analysis of evidentiary and informational value comes from T.R. Schellenberg, "The Appraisal of Modern Public Records" and "Archival Principles of Arrangement," in Daniels and Walch: 57-70 and 149-161 respectively.

[21] Gilliland-Swetland, 11.

[22] Paul Conway, *Preservation in the Digital World*, Washington, D.C.: Council on Library and Information Resources, 1996: 3. Available at http://www.clir.org/pubs/reports/conway2/index.html.

[23] Jenkinson, 11.

[24] Trudy Huskamp Peterson, "Archival Principles and Records of the New Technology," *American Archivist*, Vol. 47, no. 4., Fall 1984: 385.

[25] Mary F. Robek, Gerald F. Brown, David O. Stephens, *Information and Records Management : Document-based Information Systems, 4th edition,* New York : Glencoe (1995): 7.

[26] Research Libraries Group, "Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report," (Research Libraries Group: Mountain View, CA.), 2002: 13.

[27] Fredric M. Miller, *Arranging and Describing Archives and Manuscripts*, Chicago : Society of American Archivists, 1990: 3.

[28] The terms *data*, *file* and d*ocument* comprise components of the term record, but any further subdivision of the concept of the *record*, or a detailed analysis of semantics is beyond the scope of this paper. See Robek, Brown and Stephens, 4.

[29] Richard J. Cox, "Why Records are Important in the Information Age," *Records Management Quarterly* (January 1998): 38.

[30] James M. O'Toole, "On the Idea of Permanence," *American Archivist*, Vol. 52, Winter 1989: 11-12. O'Toole's widely cited article is the definitive exploration on the idea of permanence in archival materials. He ultimately comes to the conclusion that the more relatively applied concept of *endurance* is a more accurate description of what actually should take place in archival appraisal.

[31] Helen R. Tibbo, "On the Nature and Importance of Archiving in the Digital Age," *Advances in Computers, v. 59*, December 2003.

[32] O'Toole, "On the Idea of Permanence": 24.

[33] Jenkinson, 12.

[34] Luciana Duranti, "Reliability and authenticity: the Concepts and their Implications," Archivaria no. 39 (Spring 1995): 6-7.

[35] Duranti, 8.

[36] "Interpares Project: Authenticity Task Force Final Report," British Columbia, Canada, 2001: 3. Available at http://www.interpares.org/documents/atf_draft_final_report.pdf.

[37] This topic has seen an explosion of scholarship recently. See especially "Digicult: Integrity and Authenticity of Digital Cultural Heritage Objects," August 2002. Available at http://www.digicult.info/downloads/thematic_issue_1_final.pdf, and Council on Library and Information Resources, *Authenticity in a Digital Environment*, May 2000. Available at http://www.clir.org/pubs/reports/pub92/pub92.pdf (Accessed 10 November 2002).

[38] "Interpares Project: Authenticity Task Force Final Report," British Columbia, Canada. Available at http://www.interpares.org/documents/atf_draft_final_report.pdf .

[39] The Waters/Garrett report interchangeably uses the terms *digital object* and *information object* to refer to archival objects in electronic information systems. Each of these terms is semantically problematic, with *information object* especially fuzzy. The term *digital object* has been adopted by the intellectual property community (largely publishing interests) and applied to their digital object identifier system (see http://www.doi.org/welcome.html, accessed 9 April 2004), which makes it a little squishy. The term used in the OAIS reference model to refer to the complete collection of bits and metadata, *information package*, is also difficult to comprehend outside of the OAIS context. I acknowledge the viability of the Waters/Garrett view of the world (clarifying it slightly) and use the term *record* to refer to complex items in traditional archival systems, and the term *digital object* to refer to similar objects in digital systems.

[40] Garrett and Waters, 11.

[41] David M. Levy, "Heroic Measures: Reflections on the Possibility and Purpose of Digital Preservation," *Proceedings of the Third ACM Conference on Digital libraries*, Pittsburgh, Pennsylvania, New York: ACM Press, 1998: 153.
[42] Garrett and Waters, 14.
[43] Miller, 25-26.
[44] Anne J. Gilliland-Swetland, "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment," Washington, D.C.: CLIR, 2000: 12. Available at http://www.clir.org/pubs/reports/pub89/pub89.pdf.
[45] Ernst Posner, "Archival Development Since the French Revolution", quoted in Miller, 26.
[46] Samuel Muller, J.A. Feith, and R. Fruin, *Manual for the Arrangement and Description of Archives,* New York, The H.W. Wilson Company, 1940. As quoted in Miller, 20.
[47] The museum, art and architecture communities' definition of provenance has a representation in archival terminology in Hillary Jenkinson's concept of *custody* discussed earlier. See also Garrett and Waters, 17.
[48] Kenneth W. Duckett, *Modern Manuscripts : a Practical Manual for their Management, Care, and Use,* Nashville : American Association for State and Local History, 1975: 343.
[49] Miller, 28.
[50] Miller, 20.
[51] Miller, 31.
[52] T.R. Schellenberg, "The Appraisal of Modern Public Records," found in Maygene F. Daniels and Timothy Walch, eds., *A Modern Archives Reader : Basic Readings on Archival Theory and Practice*, Washington, D.C. : National Archives and Records Service, U.S. General Services Administration, 1984: 58.
[53] Ibid., 61.
[54] Ibid., 63.
[55] Ibid., 65.
[56] Ibid., 67.
[57] Paul Conway, *Preservation in the Digital World*, Washington, D.C.: Council on Library and Information Resources, 1996: 5. Available at http://www.clir.org/pubs/reports/conway2/index.html.
[58] Hilary Jenkinson, *A Manual of Archive Administration*, London: Percy Lund, Humphries and Co., 1966: 98.
[59] Oliver W. Holmes, "Archival Arrangement—Five Different Operations at Five Different Levels," found in Maygene F. Daniels and Timothy Walch, eds., *A Modern Archives Reader : Basic Readings on Archival Theory and Practice*, Washington, D.C. : National Archives and Records Service, U.S. General Services Administration, 1984: 162-180.
[60] Miller, 63.
[61] Miller, 62-68.
[62] Miller, 60-61.
[63] Kent M. Haworth, "Archival Description: Content and Context in Search of Structure," *Journal of Internet Cataloging*, Vol. 4, No. 3/4, 2001: 11.
[64] Miller, 22.
[65] Kent M. Haworth, "Archival Description: Content and Context in Search of Structure," *Journal of Internet Cataloging*, Vol. 4, No. 3/4, 2001: 21.
[66] The following points are derived from Anne J. Gilliland-Swetland, "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment," Washington, D.C.: CLIR, 2000: 19. Available at http://www.clir.org/pubs/reports/pub89/pub89.pdf
[67] Anne J. Gilliland-Swetland, "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment," Washington, D.C.: CLIR, 2000: v.
[68] Anne J. Gilliland-Swetland, "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment," Washington, D.C.: CLIR, 2000: 32.
[69] See Bruce I. Ambacher, ed., *Thirty Years of Electronic Records*, Lanham, MD: Scarecrow Press, 2003. This collection of essays details work at the National Archives which has addressed electronic records, including an introduction to their Electronic Records Archive (ERA). For a similar view of the Library of Congress see also *LC21: a Digital Strategy for the Library of Congress,* Washington, D.C.: National Academy Press., 2000. Available at http://books.nap.edu/catalog/9940.html.

[70] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," CCSDS 650.0-B-1. Blue Book. CCSDS: Washington, D.C. 2002. Page 4-1. Available at http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf. The analysis of the OAIS system presented in the remainder of this paper is based on this version of the document.

[71] While the OAIS model can refer to the design of both physical and digital repositories, we are largely considering it in reference to digital repositories.

[72] "The goal of this workshop was the development of US contributions to an ISO standards effort supporting the long term preservation of digital information obtained from observations of the terrestrial and space environments. The focus of this meeting was the development of proposals on the requirements, organization and content of an Archiving Reference Model for consolidation into a US contribution for the First ISO Workshop on Data Archiving held in conjunction with the Consultative Committee for Space Data Systems (CCSDS) Panel 2 Archive Standards meeting scheduled for 26-27 October, 1995 in Oxford, England." See "ISO Archiving Standards—First US Workshop" at http://ssdoo.gsfc.nasa.gov/nost/isoas/us01/ws.html.

[73] Mackenzie Smith, Associate Director for Technology, MIT Libraries. Telephone interview by author, 24 February 2004. Tape Recording.

[74] *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation*, Washington, D.C.: National Science Foundation, 2003: 10. Available at http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf.

[75] Paul Conway, Director for Information Technology Services at Duke University Libraries. Interview by author, 25 February 2004, Durham, NC. Tape Recording..

[76] Jeff Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*, Council on Library and Information Resources: Washington, D.C, 1999: 2. Available at http://www.clir.org/pubs/reports/rothenberg/pub77.pdf.

[77] Consultative Committee for Space Data Systems, iii, 1-1 - 1-3.

[78] The concept of self-validating and self-instantiating archives is taken from Bertram Ludäscher, Richard Marciano and Reagan Moore, "Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-Based Archives," Available at http://www.sdsc.edu/NARA/Publications/Web/kba.pdf.

[79] Consultative Committee for Space Data Systems, 2-1.

[80] Consultative Committee for Space Data Systems, 2-2 - 2-3

[81] Consultative Committee for Space Data Systems, 4-3 - 4-4.

[82] For more introductory material on the OAIS RM, see Brian F. Lavoie, "Meeting the Challenges of Digital Preservation: The OAIS Reference Model," OCLC Newsletter 243 (January/February 2000), p 26-30 Available at http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000001747, and Brian F. Lavoie, "The Open Archival information System Reference Model: Introductory Guide". (A joint report of the Digital Preservation Coalition (DPC) and OCLC, published electronically as a DPC Technology Report, January 2004.) Available online at: http://www.dpconline.org/docs/lavoie_OAIS.pdf.

[83] Consultative Committee for Space Data Systems, 4-8.

[84] Garrett and Waters, 13.

[85] Several projects have attempted to define a core set of preservation metadata based on the OAIS reference model. See *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*, OCLC/RLG Working Group on Preservation Metadata: Dublin, OH., 2002. Available at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf. The Current OCLC/RLG project on this front is called Preservation Metadata: Implementation Strategies (PREMIS). See its homepage at http://www.oclc.org/research/projects/pmwg/, and also Brian F. Lavoie, "Implementing Metadata in Digital Preservation Systems," *D-Lib Magazine*, April 2004. Available at http://www.dlib.org/dlib/april04/lavoie/04lavoie.html.

[86] The Harvard E-Journal archiving project was one of a set of seven Andrew W. Mellon Foundation-funded projects designed to "examine various aspects of the challenges of archiving electronic journal content." These projects were some of the first bids to explore practical digital archiving, and many attempted to implement OAIS concepts into their production systems. See http://www.diglib.org/preserve/ejp.htm.

[87] Harvard University Library, *Report on the Planning Year Grant For the Design of an E-journal Archive*, April 1, 2002. Available at http://www.diglib.org/preserve/harvardfinal.pdf.

[88] See "Cyclic Redundancy Check (CRC)" at http://www2.rad.com/networks/1994/err_con/crc.htm.

[89] See "Checksum-Wikipedia" at http://en2.wikipedia.org/wiki/Checksum (Accessed 25 November 2003).

[90] Dale Flecker, Associate Director of the Harvard University Library for Planning and Systems. Interview by author, 8 March 2004, Cambridge, Ma. Tape Recording.

[91] Mackenzie Smith, Associate Director for Technology, MIT Libraries. Telephone interview by author, 24 February 2004. Tape Recording.

[92] See Stephen Abrams and David Seaman, "Towards a Global Digital Format Registry," *Proceedings of the World Library and Information Congress: 69th IFLA General Conference and Council*, Berlin, Germany. August 1-9, 2003. Available at http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf.

[93] See "Public Records Office: PRONOM." Available at http://www.records.pro.gov.uk/pronom/.

[94] See "Request for Comments: 2046, Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types" Available at ftp://ftp.isi.edu/in-notes/rfc2046.txt.

[95] James L. Johnson, *Database: Models, Languages, Design*, New York: Oxford University Press, 1997: 3.

[96] Consultative Committee for Space Data Systems, 4-11.

[97] Garrett and Waters, 11.

[98] Consultative Committee for Space Data Systems, 4-13.

[99] Garrett and Waters, 27.

[100] Rothenberg, Jeff Rothenberg, *An Experiment in Using Emulation to Preserve Digital Publications*, Netherlands: Koninklijke Bibliotheek Den Haag, 2000: 4-5.

[101] An Open Archives Initiative service provider. See http://www.oaister.org/o/oaister/.

[102] Research Libraries Group, *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*, OCLC/RLG Working Group on Preservation Metadata: Dublin, OH, 2002: 5. Available at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

[103] *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*, OCLC/RLG Working Group on Preservation Metadata: Dublin, OH, 2002. Available at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

[104] "Metadata Encoding and Transmissions Standard (METS)." Available at http://www.loc.gov/standards/mets/.

[105] Consultative Committee for Space Data Systems, 4-21.

[106] "Unicode 4.0.0." Available at http://www.unicode.org/versions/Unicode4.0.0/.

[107] Consultative Committee for Space Data Systems, 4-27.

[108] Consultative Committee for Space Data Systems, 4-28.

[109] See "Web Naming and Addressing Overview." Available at http://www.w3.org/Addressing/.

[110] "The Handle System is a comprehensive system for assigning, managing, and resolving persistent identifiers, known as 'handles,' for digital objects and other resources on the Internet. Handles can be used as Uniform Resource Names (URNs)." See "Introduction to the Handle System." Available at http://www.handle.net/introduction.html.

[111] See "DOI System Overview - Introduction." Available at http://www.doi.org/overview/sys_overview_021601.html.

[112] Consultative Committee for Space Data Systems, 4-30.

[113] For further historical background and contextualization, see Arms, William Y., *Digital libraries*, Cambridge, Mass.: MIT Press, 2000; and Borgman, Christine L., *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*, Cambridge, Mass.: MIT Press, 2000.

[114] Robert Kahn and Robert Wilensky, "A Framework for Distributed Digital Object Services," 1995. Available at http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.

[115] William Y Arms, "Key Concepts in the Architecture of the Digital Library," *D-Lib Magazine*, July 1995. Available at http://www.dlib.org/dlib/July95/07arms.html.

[116] *LC21 : a Digital Strategy for the Library of Congress*, Washington, D.C.: National Academy Press, 2000. Available at http://books.nap.edu/catalog/9940.html.

[117] "LC 21/Digital Strategy Issued by NRC," http://www.loc.gov/today/pr/2000/00-100.html. For the complete report, see http://books.nap.edu/catalog/9940.html.

[118] "Library to Lead Nat'l Effort to Develop Digital Info Infrastructure & Preservation Program," http://www.loc.gov/today/pr/2001/01-006.html.

[119] The full text of the program and its appendices are available at http://www.digitalpreservation.gov/index.php?nav=3&subnav=1.
[120] See "Clay Shirky's Internet Writings," at http://www.shirky.com/.
[121] See "NDIIP Technical Architecture-Update," at http://www.digitalpreservation.gov/index.php?nav=3&subnav=12.
[122] See Clay Shirky, "Appendix 9. Preliminary Architecture Proposal for Long-Term Digital Preservation," in *Plan for the National Digital Information Infrastructure and Preservation Program: Appendix*, Washington, D.C.: Library of Congress, 2002: 239-240. Available at http://www.digitalpreservation.gov/repor/ndiipp_appendix.pdf.
[123] Kenneth Thibodeau, "Building the Future: The Electronic Records Archive Program," found in Ambacher, Bruce I., ed., Thirty Years of Electronic Records, Lanham, MD: Scarecrow Press, 2003: 102. NARA got only $16.3 million compared to the LC's $100-125 million.
[124] Available online at http://www.nap.edu/books/0309089476/html/.
[125] The NARA documents represent an example of how research initiatives were crossing knowledge domains and beginning to coalesce around central organizations. NARA was able to gather researchers from a broad cross-section of the information science world to review the documents relating to the development of the NARA ERA. Among the reviewers and committee were Robert Wilensky; Clifford Lynch, the director of the Coalition for Networked Information; Jerome Saltzer, Professor of Computer Science, Emeritus at M.I.T.; Margaret Hedstrom, Associate Professor at the University of Michigan School of Information; William Arms; Paul Conway, the Information Technology Services Director at the Duke University libraries; and Jeff Rothenberg.
[126] See "DLF Preservation. Preserving scholarly journals. Criteria for an archival repository 1.2." Available at http://www.diglib.org/preserve/criteria.html.
[127] Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL Bimonthly Report* 226, February 2003. Available at http://www.arl.org/newsltr/226/ir.html.
[128] Bob Boiko, *Content Management Bible*, (New York : Hungry Minds, Inc.), 2002: 111.
[129] Boiko, 81-109. Chapter seven of his book, "Introducing the Major Parts of a CMS," is an essential overview to his idea of the core functionalities of a CMS.
[130] "The company's product line 'Cumulus' is designed to manage and archive all types of digital assets used in publishing, communication, production and other workflows. The Cumulus product line offers cross-platform and Internet technology that scales completely from Enterprise to Single User." Canto promotional pamphlet, 2003.
[131] "Large-scale platforms typically marketed in multidimensional 'suites' that span many function points, but may be less well-suited for straightforward Web CMS projects. Expect about US $200-250k+ for entry-level licensing." See http://www.cmswatch.com/ContentManagement/Products/.
[132] Mackenzie Smith, Associate Director for Technology, MIT Libraries. Telephone interview by author, 24 February 2004. Tape Recording.
[133] See "Overview: Digital Repository Service (DRS)." Available at http://hul.harvard.edu/ois/systems/drs/.
[134] See "LDI Home Page." Available at http://hul.harvard.edu/ldi/.
[135] Available at http://hul.harvard.edu/ldi/resources/ldirepository.pdf.
[136] See "Functional Requirements for the LDI Repository, Phase I, Final Version." Available at http://hul.harvard.edu/ldi/resources/ldifunreq.pdf.
[137] Dale Flecker, Associate Director of the Harvard University Library for Planning and Systems. Interview by author, 8 March 2004, Cambridge, Ma. Tape Recording.
[138] See http://www.rlg.org/longterm/harvard_ejournal_archive.ppt
[139] See "Harvard University. Report on the Planning Year Grant For the Design of an E-journal Archive." Available at http://www.diglib.org/preserve/harvardfinal.html.
[140] See "Harvard University. Report on the Planning Year Grant For the Design of an E-journal Archive." Available at http://www.diglib.org/preserve/harvardfinal.html.
[141] See "Harvard University. Report on the Planning Year Grant For the Design of an E-journal Archive." Available at http://www.diglib.org/preserve/harvardfinal.html.
[142] See "JHOVE - JSTOR/Harvard Object Validation Environment." Available at http://hul.harvard.edu/jhove/jhove.html.

[143] Stephen Abrams and Marilyn Geller, "Submission Information Package (SIP) Specification: Version 1.0 Draft," Harvard University Library: Cambridge, MA, 2001: 3. Available at http://www.diglib.org/preserve/harvardsip10.pdf.
[144] Stephen Abrams and David Seaman, "Towards a Global Digital Format Registry," *Proceedings of the World Library and Information Congress: 69th IFLA General Conference and Council*, Berlin, Germany. August 1-9, 2003: 2. Available at http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf. There are a number of other resources on the subject of the development of format registries and their related data structures. See the UK Public Records Office's Pronom site at http://www.records.pro.gov.uk/pronom/; the UK's Joint Information Systems Committee (JISC) "Survey and Assessment of Sources of Information on File Formats and Software Documentation" at http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf; and John Mark Ockerbloom's "What is TOM? (And what's it good for?)" at http://tom.library.upenn.edu/intro.html.
[145] See "Harvard Archive DTD Feasibility Study," p. 4. Available at http://www.diglib.org/preserve/hadtdfs.pdf. Another example of this type of DTD development is the California Digital Library "Digital Object Standard: Metadata, Content and Encoding." Available at http://www.cdlib.org/news/pdf/CDLObjectStd-2001.pdf.
[146] This repository is still in the earliest stages of design. General information can be found at http://www.jstor.org/about/earchive.html.
[147] The theoretical and background information pages on the Dspace architecture can be found at http://www.DSpace.org/, while the live MIT Dspace site can be found at http://libraries.mit.edu/dspace-mit/.
[148] "MIT OpenCourseWare: About OCW," http://ocw.mit.edu/OcwWeb/Global/AboutOCW/about-ocw.htm.
[149] "The HP-MIT Alliance," http://www.hpl.hp.com/mit/.
[150] "MITCET overview," http://web.mit.edu/cet/overview/index.html (accessed 21 February 2003).
[151] "Dspace at MIT: Home," https://dspace.mit.edu/index.jsp.
[152] "DSpace Internal Reference Specification Technology and Architecture," Section 2.2.3. http://dspace.org/technology/architecture.pdf.
[153] "DSpace Internal Reference Specification Technology and Architecture," Section 2.2.2. http://dspace.org/technology/architecture.pdf.
[154] "DSpace Internal Reference Specification Technology and Architecture," Section 2.2.2. http://dspace.org/technology/architecture.pdf.
[155] "DSpace Internal Reference Specification Functionality," Section 2.4.4. http://dspace.org/technology/functionality.pdf.
[156] "DSpace Internal Reference Specification Functionality," Section 2.4.4. http://dspace.org/technology/functionality.pdf.
[157] "DSpace Internal Reference Specification Functionality," Section 2.4.4. http://dspace.org/technology/functionality.pdf.
[158] "DSpace Internal Reference Specification Functionality," Section 2.3.6. http://dspace.org/technology/functionality.pdf.
[159] "DSpace System Documentation: Functional Overview," http://dspace.org/technology/system-docs/functional.html.
[160] "DSpace System Documentation: Functional Overview," http://dspace.org/technology/system-docs/functional.html.
[161] "DC-Library Application Profile (DC-Lib)," http://dublincore.org/documents/2002/04/16/library-application-profile.
[162] SIMILE : Semantic Interoperability of Metadata and Information in unLike Environments. Available at http://web.mit.edu/simile/www/.
[163] "DSpace System Documentation: Functional Overview," http://dspace.org/technology/system-docs/functional.html.
[164] "DSpace Internal Reference Specification Functionality," Section 2.3.5. http://dspace.org/technology/functionality.pdf.
[165] "Resource Description Framework (RDF)," http://www.w3.org/RDF/.
[166] See "Persistent URL Home Page." Available at http://www.purl.org/.
[167] "Introduction to the Handle System," http://www.handle.net/introduction.html.

[168] There is a great deal of material being produced on this subject. For an overview, see " Free/Open Source Research Community (Home)." Available at http://opensource.mit.edu/, and Eric Raymond's "The Cathedral and the Bazaar," available at http://www.catb.org/~esr/writings/cathedral-bazaar/.

[169] " Open Source Initiative OSI - The BSD License:Licensing," http://www.opensource.org/licenses/bsd-license.php.

[170] Further information on the NARA ERA is available at http://www.archives.gov/electronic_records_archives/.

[171] Kenneth Thibodeau, "Building the Future: The Electronic Records Archives Program." Found in Bruce I. Ambacher, ed., *Thirty Years of Electronic Records*, Lanham, MD: Scarecrow Press, 2003: 92.

[172] The following analysis of the development of the ERA requirements comes from *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development*, Washington, D.C.: National Academies Press, 2003, and the *National Archives and Records Administration. National Archives and Records Administration Request for Proposal (RFP)*, 24 December 2003, NAMA-03-R-0018 for the Electronic Records Archives Amendment 0001. Available at http://www.archives.gov/electronic_records_archives/pdf/rfp.pdf and http://www.archives.gov/electronic_records_archives/pdf/requirements.pdf

[173] *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development*, Washington, D.C.: National Academies Press, 2003: 5.

[174] See "NARA | ERA | Bidders list." Available at http://www.archives.gov/electronic_records_archives/acquisition/bidders_list.html.

[175] *National Archives and Records Administration. National Archives and Records Administration Request for Proposal (RFP)*, 24 December 2003, NAMA-03-R-0018 for the Electronic Records Archives Amendment 0001: J2-9. Available at http://www.archives.gov/electronic_records_archives/pdf/rfp.pdf and http://www.archives.gov/electronic_records_archives/pdf/requirements.pdf.

[176] *National Archives and Records Administration. National Archives and Records Administration Request for Proposal (RFP)*, 24 December 2003, NAMA-03-R-0018 for the Electronic Records Archives Amendment 0001: J2-13. Available at http://www.archives.gov/electronic_records_archives/pdf/rfp.pdf and http://www.archives.gov/electronic_records_archives/pdf/requirements.pdf.

[177] The information in this paragraph is adapted from Reagan Moore, et. al., "Collection-Based Persistent Digital Archives-Part 1," *D-Lib Magazine*, Vol. 6, No. 3, March 2000. Available at http://www.dlib.org/dlib/march00/moore/03moore-pt1.html.

[178] Bertram Ludäscher, Richard Marciano and Regan Moore, "Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-Based Archives," *ACM SIGMOD Record*, 30(3), 54-63, 2001. Available at http://www.sdsc.edu/NARA/Publications/Web/kba.pdf.

[179] See "The Library of Congress Management Report on Internal Controls Over Financial Reporting: Fiscal Year Ended September 30, 1999." Available at http://www.loc.gov/fsd/fin/pdfs/fy9905.pdf.

[180] See "Metadata Encoding and Transmission Standard (METS)." Available at http://www.loc.gov/standards/mets/.

[181] These seven sections are taken from "METS: An Overview & Tutorial." Available at http://www.loc.gov/standards/mets/METSOverview.v2.html.

[182] See "METS Schemas Revised - AV Prototype." Available at http://lcweb.loc.gov/rr/mopic/avprot/metsmenu2.html.

Bibliography

The Web site addresses listed in this section were valid as of as May 2, 2004.

Abrams, Stephen, and David Seaman. "Towards a Global Digital Format Registry." Proceedings of the World Library and Information Congress: 69th IFLA General Conference and Council, Berlin, Germany. August 1-9, 2003. Available at http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf.

Abrams, Stephen, and Marilyn Geller. "Submission Information Package (SIP) Specification: Version 1.0 Draft." Harvard University Library: Cambridge, MA. 2001. Available at http://www.diglib.org/preserve/harvardsip10.pdf.

Ambacher, Bruce I., Information Technology Specialist, Modern Archives Program, National Archives and Records Administration. Telephone interview by author, 30 March 2004. Tape Recording.

Ambacher, Bruce I., ed. *Thirty Years of Electronic Records*. Lanham, MD: Scarecrow Press. 2003.

Archival Information Package (AIP) Design Study. LC-DAVRS-07. Washington, D.C.: Library of Congress. 2001. Available at http://www.loc.gov/rr/mopic/avprot/AIP-Study_v19.pdf.

Arms, William Y. *Digital libraries*. Cambridge, Mass.: MIT Press. 2000.

________. "Key Concepts in the Architecture of the Digital Library." *D-Lib Magazine*. July 1995. Available at http://www.dlib.org/dlib/July95/07arms.html.

Authenticity in a Digital Environment. Washington, D.C.: Council on Library and Information Resources. 2000. Available at http://www.clir.org/pubs/reports/pub92/pub92.pdf.

Bearman, David. *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh: Archives & Museum Informatics. 1994.

Boiko, Bob. *Content Management ible*. New York : Wiley Publishing. 2002.

Borgman, Christine L. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, Mass.: MIT Press. 2000.

Browning, Paul, and Mike Lowndes. "JISC TechWatch Report: Content Management Systems." TSW 01-02. Joint Information Systems Committee: London. 2001. Available at http://www.jisc.ac.uk/index.cfm?name=techwatch_report_0102.

Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving. Washington D.C.: Council on Library and Information Resources, 2002. Available at http://www.clir.org/pubs/reports/pub106/pub106.pdf.

Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development. Washington, D.C.: National Academies Press. 2003. Available at http://www.nap.edu/books/0309089476/html/.

Coleman, Patrick T. "Document Management for the Masses." *SW Expert*, November 1999. 50-57.

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

Conway, Paul, Director for Information Technology Services at Duke University Libraries. Interview by author, 25 February 2004, Durham, NC. Tape Recording.

________. *Preservation in the Digital World*. Washington, D.C.: Council on Library and Information Resources. 1996. Available at http://www.clir.org/pubs/reports/conway2/index.html.

Cox, Richard J. "Electronic Systems and Records Management in the Information Age: An Introduction." *Bulletin of the American Society for Information Science,* June/July 1997.

________. "Why Records are Important in the Information Age." Records Management Quarterly (January 1998).

Daniels, Maygene F., and Timothy Walch, eds. A Modern Archives Reader : Basic Readings on Archival Theory and Practice. Washington, D.C. : National Archives and Records Service, U.S. General Services Administration. 1984.

Duckett, Kenneth W. *Modern Manuscripts : a Practical Manual for their Management, Care, and Use.* Nashville : American Association for State and Local History. 1975.

Duranti, Luciana. "Reliability and authenticity: the Concepts and their Implications." *Archivaria* no. 39 (Spring 1995).

E-Journal Archive DTD Feasibility Study. Inera Inc.: Boston. 2001. Available at http://www.diglib.org/preserve/hadtdfs.pdf.

Flecker, Dale, Associate Director of the Harvard University Library for Planning and Systems. Interview by author, 8 March 2004, Cambridge, Ma. Tape Recording.

Fleischhauer, Carl, Project Coordinator, Office of Strategic Initiatives, Library of Congress. Telephone interview by author, 29 March 2004. Tape Recording.

Galloway, Patricia Kay, Assistant Professor, Archival Enterprise and Digital Asset Management, School of Information, University of Texas-Austin. Interview by author, 27 February 2004, Austin, Tx. Tape Recording

Garrett, John and Donald Waters, eds. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Mountain View, CA: Research Libraries Group. May 1996. Available at http://www.rlg.org/ArchTF/.

Gilliland-Swetland, Anne J. "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment." Washington, D.C.: Council on Library and Information Resources. February 2000. Available at http://www.clir.org/pubs/reports/pub89/pub89.pdf.

Greenan, Monica. *ERPANET OAIS Training Seminar Report*. Glasgow, United Kingdom: Erpanet. 2003. Available at http://www.erpanet.org/www/products/copenhagen/ERPANET%20OAIS%20Training%20Seminar%20Report_final.pdf.

Haworth, Kent M. "Archival Description: Content and Context in Search of Structure." *Journal of Internet Cataloging*, Vol. 4, No. 3/4, 2001: 7-26.

Hedstrom, Margaret, et al. "It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation." Washington D.C.: National Science Foundation. August 12, 2002. Available at http://www.si.umich.edu/digarch/Report.DFt.2.doc.

Henson, Steve. *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries.* Washington D.C.: Library of Congress. 1983.

Hilton, James. "New Horizons: Digital Asset Management Systems." *Educause Review*. March/April 2003.

Interpares Project: Authenticity Task Force Final Report. British Columbia, Canada. 2001. Available at http://www.interpares.org/documents/atf_draft_final_report.pdf.

Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation. Washington, D.C.: National Science Foundation. 2003. Available at http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf.

Jenkinson, Hilary. *A Manual of Archive Administration*. London: Percy Lund, Humphries and Co. 1966

Johnson, James L. *Database: Models, Languages, Design*. New York: Oxford University Press. 1997.

Kahn, Robert and Robert Wilensky. "A Framework for Distributed Digital Object Services." 1995. Available at http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.

Lagoze, Carl. "A Secure Repository Design for Digital Libraries." *D-Lib Magazine*. December 1995. Available at http://www.dlib.org/dlib/december95/12lagoze.html.

Lagoze, Carl, ed. "Core Services in the Architecture of the National Digital Library for Science Education (NSDL)." Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital libraries. Portland, Oregon. 2002. New York: ACM Press: 201 - 209.

Lavoie, Brian F. "Implementing Metadata in Digital Preservation Systems." *D-Lib Magazine*. April 2004. Available at http://www.dlib.org/dlib/april04/lavoie/04lavoie.html.

________. "Meeting the Challenges of Digital Preservation: The OAIS Reference Model." *OCLC Newsletter* January/February 2000: 26-30.

________. "The Incentive to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making." Dublin, OH.: OCLC. April 2003. Available at http://www.oclc.org/research/projects/digipres/incentives-dp.pdf.

________. "The Open Archival information System Reference Model: Introductory Guide." Dublin, OH: Digital Preservation Coalition/OCLC. 2004. Available online at: http://www.dpconline.org/docs/lavoie_OAIS.pdf.

LC21 : a Digital Strategy for the Library of Congress. Washington, D.C. : National Academy Press. 2000. Available at http://books.nap.edu/catalog/9940.html.

Levy, David M. "Heroic Measures: Reflections on the Possibility and Purpose of Digital Preservation." Proceedings of the Third ACM Conference on Digital libraries, Pittsburgh, Pennsylvania. New York: ACM Press. 1998.

Library of Congress Digital Audio-Visual Repository System (DAVRS) Conceptual Design Document. Prepared by User Technology Associates and QB Incorporated. LC-DAVRS-03. Washington, D.C.: Library of Congress. 2000.

Library of Congress Digital Audio-Visual Repository System (DAVRS) System Requirements Document. LC-DAVRS-02. Washington, D.C.: Library of Congress. 2000.

Lorie, Raymond A. "A Project on Preservation of Digital Objects." *RLG Diginews*, Vol. 5, No. 3, June 15, 2001. Available at http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2.

________. "Long-Term Archiving of Digital Information." IBM Research: Yorktown Heights, NY. May 2000. Available at http://domino.watson.ibm.com/library/cyberdig.nsf/Home.

Ludäscher, Bertram, Richard Marciano and Regan Moore. "Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-Based Archives." *ACM SIGMOD Record*, 30(3), 54-63. 2001. Available at http://www.sdsc.edu/NARA/Publications/Web/kba.pdf.

Lynch, Clifford A. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." *ARL Bimonthly Report* 226, February 2003. Available at http://www.arl.org/newsltr/226/ir.html.

_______. "The Integrity of Digital Information: Mechanics and Definitional Issues." *Journal of the American Society for Information Science*, 45(10): 737-744, 1994.

________. "When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web." *Journal of the American Society for Information Science and* Technology, 52(1): 12-17. 2001.

McCord, Alan. "Overview of Digital Asset Management Systems." Educause Evolving Technologies Committee: Washington, D.C. September 6, 2002. Available at http://www.educause.edu/ir/library/pdf/DEC0203.pdf.

Mellon Fedora Technical Specification. University of Virginia: Charlottesville, VA. 2002. Available at http://www.fedora.info/documents/master-spec-12.20.02.pdf.

Miller, Fredric M. *Arranging and Describing Archives and Manuscripts*. Chicago : Society of American Archivists. 1990.

Moore, Reagan. "Common Consistency Requirements for Data Grids, Digital Libraries, and Persistent Archives." San Diego: San Diego Supercomputer Center. 2001. Available at http://www.ggf.org/Meetings/ggf7/drafts/GGF7_Data_Consistency.Word95.pdf.

________. "The San Diego Project: Persistent Objects." San Diego: San Diego Supercomputer Center. No date. Available at http://www.sdsc.edu/NARA/Publications/persistent-objects.doc.

________, Arcot Rajasekar, and Richard Marciano. "Collection-Based Persistent Archives." San Diego, Ca.: San Diego Supercomputer Center. No date. Available at http://www.sdsc.edu/NARA/Publications/OTHER/Persistent/Persistent.html.

________, Richard Marciano and Bertram Ludäscher. "The Senate Legislative Activities Collection (SLA): a Case Study: Infrastructure Research to Support Preservation Strategies." San Diego: San Diego Supercomputer Center. January 18, 2001. Available at http://www.sdsc.edu/TR/TR-2001-05.doc.pdf.

________ et al. "Collection-Based Long-Term Preservation." San Diego: San Diego Supercomputer Center. No date. Available at http://www.sdsc.edu/NARA/Publications/nara.pdf.

________ et al. "Collection-Based Persistent Digital Archives-Part 1." *D-Lib Magazine,* Vol. 6, No. 3. March 2000. Available at http://www.dlib.org/dlib/march00/moore/03moore-pt1.html.

________ et al. "Collection-Based Persistent Digital Archives-Part 2." *D-Lib Magazine*, Vol. 6, No. 4. April 2000. Available at http://www.dlib.org/dlib/april00/moore/04moore-pt2.html.

________ et al. "Configuring and Tuning Archival Storage Systems." San Diego: San Diego Supercomputer Center No date. Available at http://www.sdsc.edu/NARA/Publications/OTHER/HPSS-tuning/HPSS-tun.v3.html.

Muller, Samuel, J.A. Feith, and R. Fruin. *Manual for the Arrangement and Description of Archives.* New York : The H.W. Wilson Company. 1940.

National Archives and Records Administration Request for Proposal (RFP). 24 December 2003. NAMA-03-R-0018 for the Electronic Records Archives Amendment 0001. Washington, D.C.: National Archives and Records Administration. Available at http://www.archives.gov/electronic_records_archives/pdf/rfp.pdf.

Nelson, Michael L. and Kurt Maly. "Buckets: Smart Objects for Digital Libraries." *Communications of the ACM*, Vol. 44, No. 5. May 2001: 60-62.

O'Toole, James M. "On the Idea of Permanence." *American Archivist*, Vol. 52, Winter 1989.

________. *Understanding Archives and Manuscripts*. Chicago: Society of American Archivists. 1990.

Payette, Sandra and Thornton Staples. "The Mellon Fedora Project: Digital Library Architecture Meets XML and Web Services," Sixth European Conference on Research and Advanced Technology for Digital Libraries. *Lecture Notes in Computer Science*, Vol. 2459. New York: Springer-Verlag. 2002: 406-421. Available at http://www.fedora.info/documents/ecdl2002final.pdf.

Peccia, Nestor. "CCSDS Standards: A Reference Model for an Open Archival Information System (OAIS)." European Space Operations Center: Darmstadt, Germany. 1998. Available at http://www.nasda.go.jp/pr/event/app/spaceops/paper98/track3/3b002.pdf.

Peterson, Trudy Huskamp. "Archival Principles and Records of the New Technology." *American Archivist*, Vol. 47, no. 4. Fall 1984.

Plan for the National Digital Information Infrastructure and Preservation Program. Washington, D.C.: Library of Congress. 2002. Available at http://www.digitalpreservation.gov/repor/ndiipp_plan.pdf.

Plan for the National Digital Information Infrastructure and Preservation Program: Appendix. Washington, D.C.: Library of Congress. 2002. Available at http://www.digitalpreservation.gov/repor/ndiipp_appendix.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: Dublin, OH. 2002. Available at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

Robek, Mary F., Gerald F. Brown, David O. Stephens. *Information and Records Management : Document-based Information Systems, 4th edition*. New York : Glencoe. 1995.

Rothenberg, Jeff. *An Experiment in Using Emulation to Preserve Digital Publications*. Netherlands: Koninklijke Bibliotheek Den Haag. 2000. Available at http://www.kb.nl/coop/nedlib/results/emulationpreservationreport.pdf.

________. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Council on Library and Information Resources: Washington, D.C. 1999. Available at http://www.clir.org/pubs/reports/rothenberg/pub77.pdf.

________. *Ensuring the Longevity of Digital Information*. Santa Monica, CA: Rand Corporation. February 22, 1999. Available at http://www.clir.org/pubs/archives/ensuring.pdf.

Smart, L.J. "OAIS, METS, MPEG-21, and Archival Values." *The Moving Image: The Journal of the Association of Moving Image Archivists*. Vol. 2, No. 1. Spring 2002.

Smith, Mackenzie, Associate Director for Technology, MIT Libraries. Telephone interview by author, 24 February 2004. Tape Recording.

________, Associate Director for Technology, MIT Libraries. Interview by author, 15 March 2004, Cambridge, Ma. Tape Recording

Staples, Thornton and Ross Wayland. "Virginia Dons FEDORA." *D-Lib Magazine*, Vol. 6, No. 7/8. July/August 2000. Available at http://www.dlib.org/dlib/july00/staples/07staples.html.

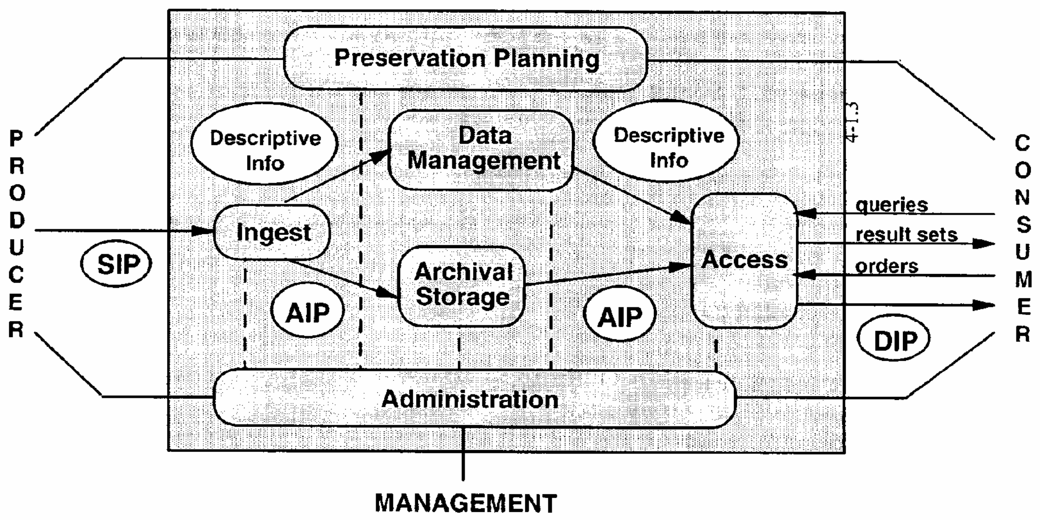
________, Ross Wayland, and Sandra Payette. "The Fedora Project: An Open Source Digital Repository Management System." *D-Lib Magazine*, Vol. 9, No. 4. April 2003. Available at http://www.dlib.org/dlib/april03/staples/04staples.html.

Tibbo, Helen R.. "On the Nature and Importance of Archiving in the Digital Age." *Advances in Computers.* v. 57, 2003: 1-67.

Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report. Mountain View, CA: Research Libraries Group. 2002. Available at http://www.rlg.org/longterm/repositories.pdf.

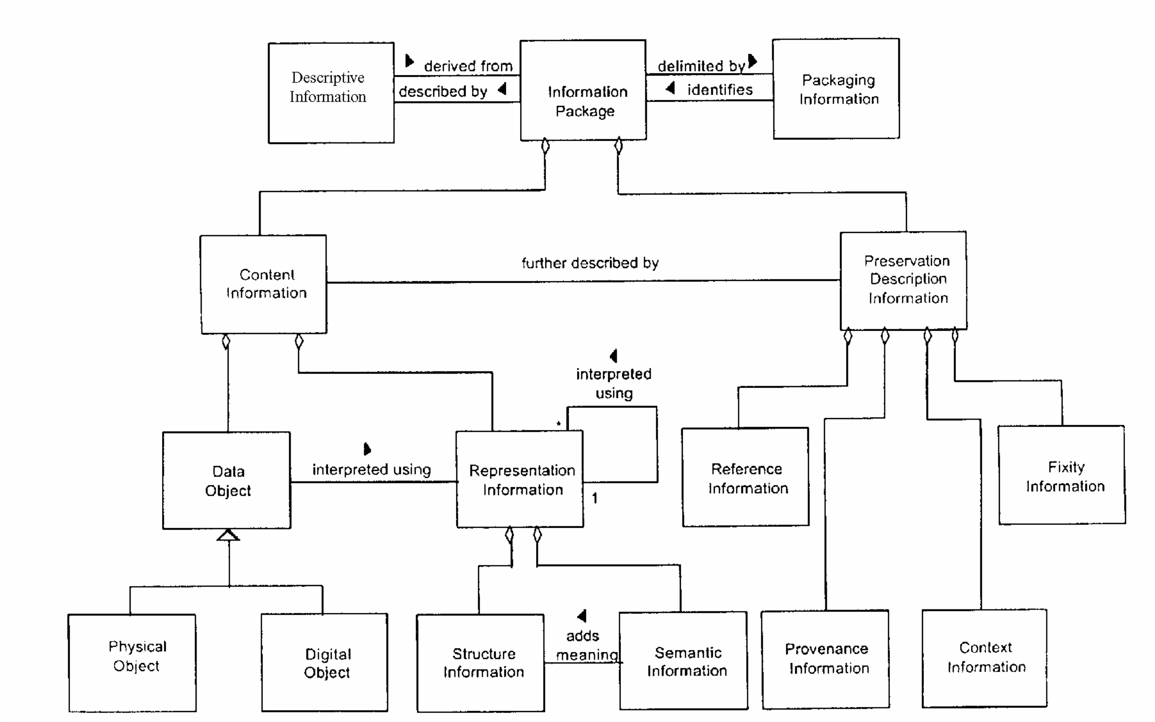
Werf, Titia van der. "The Deposit System for Electronic Publications: A Process Model." Nedlib Report Series 6. Nedlib Consortium: The Hague, Netherlands. 2000. Available at http://www.kb.nl/coop/nedlib/results/DSEPprocessmodel.pdf.

Appendix A.1: The OAIS Functional Model



OAIS Functional Entities

Appendix A.2: The OAIS Information Model



The Archival Information Package (AIP), detailed view

Appendix A.3: Academic Affairs Institutional Review Board Proposal Outline

THE UNIVERSITY OF NORTH CAROLINA
AT
CHAPEL HILL

Student Research Project
School of Information and Library Science
Phone# (919) 962-8366
Fax# (919) 962-8071

The University of North Carolina at Chapel Hill
CB# 3360, 100 Manning Hall
Chapel Hill, N.C. 27599-3360
info@ils.unc.edu

Academic Affairs Institutional Review Board Proposal Outline

Study: *The Ghost in the Machine: Traditional Archival Practice in the Design of Digital Repositories for Long-Term Preservation*
Principal Investigator: William Lazorchak
Advisor: Dr. Helen Tibbo, SILS, University of North Carolina at Chapel Hill

- **Project Description**

One of the core duties of the archivist is to select materials and provide for their long-term preservation. The materials in question have traditionally been physical items such as papers, photographs and related ephemera, but increasingly the materials in question are born-digital from computer information systems. Administering the long-term preservation of physical items has been challenging, but the preservation of digital information is proving to be even more complex.

Digital preservation has been defined as the managed activities necessary for ensuring both the long-term maintenance of a bytestream of digital data and the continued accessibility of its contents.[1] These *managed activities* of the digital preservation imperative have made significant demands on archivists, many of whom are technically and fiscally ill-equipped to deal with the issue. The research community of academics and university alliances, standards-making bodies such as the World Wide Web Consortium (W3C), open source software developers and forward-thinking commercial developers is working in widely disparate areas to attempt to gain some control over the vexations related to the management and preservation of digital materials.

[1] Research Libraries Group, "Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report," (Research Libraries Group: Mountain View, CA.), 2002: 3

The development of trusted repositories in the digital world which echo the physical spaces of traditional archives is a core area of digital archiving research, but the process of building this deep infrastructure in the world of digital preservation is only just beginning.[2] Despite several decades of awareness of the acute digital preservation needs, the advances in systems designed for digital preservation have been marked by small, tentative steps. For example, a coherent definition of the attributes and responsibilities of digital repositories which exhibit trust metrics has only recently been articulated.[3]

The conception of a trusted digital repository is actualized in an electronic information system as a software architecture and its associated management responsibilities which aggregates an array of tools, hardware components and business rules into a coherent system capable of providing the necessitated range of preservation-oriented functions. These repository systems would effectively replicate the functions found in a traditional brick-and-mortar archive to the extent that those functions were necessary in the digital realm.

The research community has begun to coalesce around the high-level design provided by the Reference Model for an Open Archival Information System (OAIS)[4], a dense 150 page document which defines the functional requirements of a physical or digital archival system which purports to address long-term preservation requirements. Despite its many merits, the reference model provides only a conceptual framework, remaining agnostic to possible implementations, and only hinting at possible answers from its high-level perch.

This research undertakes an analysis of the application of the OAIS Reference Model to the development of repository architectures designed for long-term preservation. I am researching how the OAIS model is translated into a physical, functioning system architecture, and I also have a strong interest in identifying how traditional archival values such as provenance, original order, and the certification of file authenticity and reliability are represented in an OAIS-informed repository design. While physical architecture design and archival mappings are the primary focus of my research, I'm interested in the complete range of issues surrounding OAIS implementations which effect its success (or failure) as a model.

The body of literature detailing implementations of the OAIS model is still quite small. I have immersed myself in this literature for the past two years, but have determined that the best way to gain real knowledge about implementation strategies is to approach the handful of researchers and institutions whose work has been applied to the OAIS model and to query them directly on their findings. My desire is to conduct interviews with the principles involved in the research projects, either in person, on the telephone, or through email. I am especially interested in making site visits to the

[2] John Garrett and Donald Waters, editors, "Preserving Digital Information -- Report of the Task Force on Archiving of Digital Information," May 1996. Available at http://www.rlg.org/ArchTF/. (Accessed 5 May 2003): 7.

[3] Research Libraries Group, *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report*, Research Libraries Group: Mountain View, CA, 2002. Available at http://www.rlg.org/longterm/repositories.pdf. (Accessed 22 October 2002).

[4] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," CCSDS 650.0-B-1. Blue Book. CCSDS: Washington, D.C. 2002. Page 4-1. Available at http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf. (Accessed 7 October 2002).

institutions which have attempted implementations of the OAIS model to view the implementation in the context of its development.

- **Participants**

a) Age, sex and approximate number.

The age of the participants is of no consequence to the research, and will not be tabulated. However, it can be concluded that all participants will be over the age of eighteen due to the fact that the potential interviewees hold upper-level management positions in academic libraries. The sex can be determined by the participant's names, but is of no consequence to the research and will not be documented. Thirty potential participants have been identified.

b) Inclusion/exclusion criteria

Participants have been selected based on their participation in research projects directly related to an attempt to implement the OAIS reference model. Their names have been gathered from publicly available documents supporting these research efforts.

c) Method of recruiting

Participants have been identified based on the above criteria, and their contact information has been harvested from publicly available sources. Initial contact will be made through electronic mail.

d) Inducement of participation

No inducement to participate is offered

- **Are participants at risk?**

There should be no physical or emotional risk to the potential participants.

[Question #4 is not applicable]

- Are Illegal activities involved?

  No

- Is deception involved?

  No

- **Potential Benefit to the Participants and the Research Community**

Benefit potential to the participant includes the further dissemination of their research findings. This dissemination will benefit the research community by making information more widely available.

- **How will prior consent be obtained?**

If the interview with the participant is being conducted in person, the participant will be given a copy of the consent form prior to the interview and requested to sign it. If the interview is being conducted by telephone or email, a copy of the consent form will be sent to the participant by electronic mail for review. If the participant gives their consent to the interview, a hard copy of the form will be mailed to them in a self-addressed stamped envelope for them to sign and return.

If any participant is uncomfortable with the terms of consent they are free to abstain from the interview, or to request that their interview be used for background purposes only. A copy of the consent form is attached.

- **Describe security procedures for privacy and confidentiality**

Unless privacy constraints are specifically requested by the participants, all of the information gathered will be made publicly available and participants will be cited.

Appendix A.4: Sample Participant Solicitation Letter

THE UNIVERSITY OF NORTH CAROLINA
AT
CHAPEL HILL

Student Research Project
School of Information and Library Science
Phone# (919) 962-8366
Fax# (919) 962-8071

The University of North Carolina at Chapel Hill
CB# 3360, 100 Manning Hall
Chapel Hill, N.C. 27599-3360
info@ils.unc.edu

Sample Participant Solicitation Letter

Study: *The Ghost in the Machine: Traditional Archival Practice in the Design of Digital Repositories for Long-Term Preservation*
Principal Investigator: William "Butch" Lazorchak
Advisor: Dr. Helen Tibbo, SILS, University of North Carolina at Chapel Hill

****

Dear Colleague,

My name is Butch Lazorchak, and I am a Master's candidate in the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill. I am researching and writing my Master's thesis which is exploring the application of traditional archival principles to the design of digital repositories for long-term preservation. For example, I'm interested in identifying how traditional archival values such as provenance, original order, and the certification of file authenticity and reliability are represented in the designs of functioning digital repositories. My early understanding of how these concepts are represented is fully informed by the Open Archival Information Systems reference model (OAIS).

While physical architecture designs and their associated archival mappings are the primary focus of my research, I'm also interested in the complete range of issues surrounding digital repository implementations which might influence their successful uptake, including the social, political and economic issues.

I have immersed myself in the literature discussing digital repository architectures for the past two years, but have determined that the best way to gain real knowledge about implementation strategies is to approach the handful of researchers and institutions who are working on repository architectures and query them directly on their findings.

You and your organization are doing advanced research on the systems that I'm examining, and I'm hoping that you'll be able to spare some time to speak with me about your experiences. My desire is to conduct interviews with the principles involved in the research projects, either in person, on the telephone, or through email. I am especially interested in making site visits to institutions on the east coast which have functioning implementations of a digital repositories to view the implementation in the context of its development. The material gathered in these interviews would be used as background information or be cited in my final thesis.

My work is being carried out under the tutelage of my faculty advisor Dr. Helen Tibbo, and in consultation with Dr. Gary Marchionini. Dr. Tibbo can be reached at (919) 962-8063 and tibbo@ils.unc.edu. Dr. Marchionini can be reached at (919) 966-3611 and march@ils.unc.edu. The project has received approval from the Academic Affairs Institutional Review Board at UNC-CH. I have composed a set of example questions that I will gladly supply in advance.

I know that your time is valuable, but your assistance would be greatly appreciated. If you have questions regarding this research, I encourage you to contact me at home at (919) 489-4799, cell phone (919) 423-4425 or by email at Butch@squealermusic.com. Thank you in advance for your consideration of this project, and I hope to hear from you soon.

Sincerely,

Butch Lazorchak
Masters candidate, School of Information and Library Science
University of North Carolina-Chapel Hill
http://www.butchlazorchak.org

Appendix A.5: Consent to Participate in Research

THE UNIVERSITY OF NORTH CAROLINA
AT
CHAPEL HILL

Student Research Project
School of Information and Library Science
Phone# (919) 962-8366
Fax# (919) 962-8071

The University of North Carolina at Chapel Hill
CB# 3360, 100 Manning Hall
Chapel Hill, N.C. 27599-3360
info@ils.unc.edu

Consent to Participate in Research

Study: *The Ghost in the Machine: Traditional Archival Practice in the Design of Digital Repositories for Long-Term Preservation*
Principal Investigator: William Lazorchak
Advisor: Dr. Helen Tibbo, SILS, University of North Carolina at Chapel Hill

• Introduction

You are asked to participate in a research study conducted by William Lazorchak, a Master's candidate from the School of Information and Library Science at the University of North Carolina at Chapel Hill. The results of this research will contribute to Mr. Lazorchak's Master's paper. You have been selected as a possible participant in the study because you have been a contributor to one of the projects identified as having great significance in addressing the issues surrounding the implementation of the Open Archival Information System (OAIS) reference model.

• Purpose of the Research

This research undertakes an analysis of the application of the OAIS Reference Model to the development of repository architectures designed for long-term preservation. I am researching how the OAIS model is translated into a physical, functioning system architecture, and I also have a strong interest in identifying how traditional archival values such as provenance, original order, and the certification of file authenticity and reliability are represented in an OAIS-informed repository design. While physical architecture design and archival mappings are the primary focus of my research, I'm interested in the complete range of issues surrounding OAIS implementations which effect its success (or failure) as a model.

• Procedures

The body of literature detailing implementations of the OAIS model is still quite small. I have immersed myself in this literature for the past two years, but have determined that the best way to gain real knowledge about implementation strategies is to approach the handful of researchers and institutions whose work has been applied to the OAIS model and to query them directly on their findings. There is no remuneration involved in participating in this research. Participants will be interviewed, either in person, over the phone or by email. A prospective interview protocol is available to potential participants in advance. The interview will take approximately one hour.

• Potential Risks and Discomforts

There should be no physical or emotional risk to the potential participants.

• Potential Benefits to Participants and/or to Society

Benefit potential to the participant includes the further dissemination of their research findings. This dissemination will benefit the research community by making information more widely available.

• Confidentiality

There are a small number of institutions participating in implementations of the OAIS, making the identification of sources relatively easy to detect. In the interest of full disclosure and knowledge sharing, we request that participants allow their responses to be cited. Participants may request a copy of the final report for review prior to publication. Audio tapes of the interviews will be retained until transcriptions have been made. Upon completion and review of the transcriptions the audio tapes will be erased.

• Participation and Withdrawal

You can choose whether to participate in this study or not, and you may withdraw at any time without penalty. You may also refuse to answer any questions you do not want to answer and still remain in the study. The investigator may withdraw you from this research if circumstances arise which warrant doing so. Possible circumstances might include an overabundance of material, or a preponderance of material not specifically appropriate to the topic.

• Identification of Investigators and Review Board

If you have any questions or concerns about the research, please feel free to contact:

Principal Investigator:
William "Butch" Lazorchak
2701 Old Sugar Road
Durham, NC 27707
(919) 489-4799
contactus@squealermusic.com

Faculty Sponsor:
Dr. Helen R. Tibbo
Professor, School of Information and Library Science

Room 201 Manning Hall
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
(919) 962-8063
tibbo@ils.unc.edu

If you have questions or concerns about your rights as a participant, you may contact the Academic Affairs Institutional Review Board at (919) 962-7761 or aa-irb@unc.edu.

Signature of Research Participant

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this research and to allow my responses to be cited. I have been provided a copy of this form.

Name of Participant (please print):____

Signature of Participant:____

Date:______________

Appendix A.6: Prospective Interview Protocol Questions

THE UNIVERSITY OF NORTH CAROLINA
AT
CHAPEL HILL

Student Research Project
School of Information and Library Science
Phone# (919) 962-8366
Fax# (919) 962-8071

The University of North Carolina at Chapel Hill
CB# 3360, 100 Manning Hall
Chapel Hill, N.C. 27599-3360
info@ils.unc.edu

Prospective Interview Protocol Questions

Study: *The Ghost in the Machine: Traditional Archival Practice in the Design of Digital Repositories for Long-Term Preservation*
Principal Investigator: William Lazorchak
Advisor: Dr. Helen Tibbo, SILS, University of North Carolina at Chapel Hill

****

Personal Information:

Name:
Company/Affiliation:
Job Title:

Describe the responsibilities of your current position:

What background do you bring to your current position? Can you briefly run down your previous employment and education experiences that you've been able to apply to your current position?

General Background on Digital Preservation Issues:

How active has your organization been in the past in confronting digital preservation issues? Can you name some of the more significant recent projects directly related to digital preservation issues?

What has proven to be the most difficult problem facing institutions attempting to archive digital materials?

One of the more politically charged issues relating to the preservation of digital resources is what sort of responsibility libraries have to archive materials that they may not actually own. This becomes especially important when considering electronic journals. What responsibilities do libraries have to archive digital materials such as ejournals to ensure that they are preserved in some way?

One of the more significant question in the early stages of digital repository design is whether the archive in question should be dark or light. "Dark archives" are those that house material which is not publicly available, light is the opposite. What is the initial status of your organization's archive on the dark-light continuum?

The dichotomy of darkness/lightness introduces the concepts of "trigger events" (events that would cause previously dark materials to become light) and the "moving wall," a predefined schedule of temporal trigger events. Can you iterate any discussions your organization may have had in relation to trigger events and their associated moving walls?

These trigger events largely come into play when considering proprietary content, such as that owned by electronic journal publishers. If you're able to speak on the issue, what would be appropriate trigger events from the ejournal publisher's perspective that would open up the material in the archival repositories? Are publishers averse to predetermining a temporal "moving wall?"

What constituencies did you query about the possible uses of the archived materials? Archivists are familiar with the concept of secondary use of historical materials. Is it possible that other constituencies should be queried about the attributes of electronic archived materials that need to be preserved?

Background on the Open Archival Information System Reference Model:

Are you familiar with the report from John Garrett and Donald Waters called Preserving digital information: Report of the Task Force on Archiving of Digital Information from May 1996? It is accessible at http://www.rlg.org/ArchTF/.

If you are familiar with the report, can you comment on whether the advancement of the preservation of digital materials has improved or changed much since the publication of the report in 1996?

What influenced your organization's decision to pursue the construction of a repository, as opposed to approaching the digital preservation issue from another angle?

How did your institution become interested in the Open Archival Information System (OAIS) Reference Model?

How familiar were you personally with the OAIS reference model prior to the beginning of the project?

Is the OAIS model too abstract to provide much implementation guidance?

What kind of digital repository infrastructure was already in place at your organization prior to the consideration of the OAIS model?

What efforts had your institution made in the past that allowed you to feel that it was prepared to undertake an OAIS repository project?

What academic structures are leading the development of the OAIS model? That is, do you believe that the interest in the OAIS model is originating chiefly from the library/archives sector, the computer science sector or the business/economics sector?

Speaking very generally, how are OAIS repository architectures different than the commercial content management architectures being developed by companies such as Documentum, Vignette, Artesia, Avid, etc.?

Speaking very generally, how are OAIS repository architectures different than electronic records management (ERMS) systems?

If there are similarities (or perhaps even a great deal of crossover) between OAIS architectures and the above types of systems, what differentiates an OAIS-modeled system and makes it an important topic for research?

Implementation of the OAIS for your Specific Project:

Discuss how your organization began its consideration of the OAIS model in the early stages.

After considering the OAIS, how did your organization come to a determination about what aspects of the model to focus on when proceeding with an implementation plan?

Several projects have made the decision to focus on the preservation of semantic-level information ("source forms"), as opposed to syntactic-level information ("presentation forms"). The debate between these two approaches in the archival community has been fairly inconclusive. Is semantic preservation the only possible alternative when considering digital information? Do archivists hold syntactic representation too dearly? Where does the dividing line on this issue seem to lay? Is it possible that both forms should be represented in the archive? Describe some of the debate that led your organization to decide to go one way or the other...

Many of the most active OAIS projects are considering ejournals, so many of my questions will be framed in those terms. Several projects have noted that ejournal items might best be catalogued at the *article* level as opposed to the *issue* level, especially for an archival repository that was subject-specific. How has your organization considered this for the material it is ingesting into an archival repository? Is the concept of *issue* still an appropriate one for these types of information?

One key point noted by several projects was that "smart automation" of repository processes could reduce labor and save costs. In the case of several of the ejournal archiving projects, that "smart automation" seems to have been achieved by a strict requirement of publisher compliance to formal standards for Submission Information Packages (SIPS) and by the definition of a small set of normative data formats acceptable for archiving. In these early stage of experimental development, is most of the burden of standardizing ingest items moving upstream to the creators? Are creators prepared (or willing) to handle that burden? Will repositories continue to restrict their input to a normative set of data formats, or do you foresee that technological developments will open up the number of formats capable of being successfully preserved?

The above topic touches on the "migration vs. emulation" debate. What position does your organization take on migration or emulation as potential solutions to the need to continually refresh fragile pieces of data?

Several projects have determined that the Metadata Encoding and Transmission Standard (METS) coming out of the Library of Congress would be the most viable metadata format to use for the Submission Information Packages (SIPS). What other metadata formats were considered for the repository architecture, and what were/are their relative strengths and weaknesses?

Discuss the choices your organization had to make in the early stages of the project in determining the actual content that would be preserved in the repository. What precisely is stored in the archive (what types of information, what formats, etc.)?

In the case of ejournals, this would included the decisions about which parts of the ejournal to include (frontmatter, threaded discussions=yes, advertisements=no).

How is access to the material in the repository guaranteed?

Who owns the material in the repository?

How and under what circumstances is it accessed?

Who authorizes that access?

Will the potential repository be a "fail-safe" repository, or will it provide the same functionality as the original digital items in their original form? Are both options possible?

To what degree has your organization relied on open source software tools to build the repository? Or has the repository architecture been designed to be compatible with an existing proprietary system?

Describe the tools used to build the repository architecture (including software components, etc.)? How much programming had to take place to get the existing tools to fit your organization's conception of the finished product? Describe some of the programming solutions that took place at your organization.

What findings has your project made regarding ingest validation at a semantic level? Is this a huge issue that will take up significant resources, or is human semantic validation only necessary for high-level oversight, with little significant overhead to the operating budget added?

How "clean" are the XML files coming from publishers and/or submitters? Have publishers fully embraced XML encoding to the point where the files were close to being totally ready for ingestion, or was there quite a bit of negotiation with the publishers to get their files in an acceptable form? What steps need to be taken to ensure that information arrives at the archive in reasonably good form?

Has there been any conflict between the XML entity sets of different publishers or submitting organizations? If so, how has that been reconciled?

Will information about the items ingested into the digital repository be stored in the library's OPAC? If your organization has chosen to have the materials remain dark, how will they be represented in your OPAC? How is information transmitted from the INGEST function of the OAIS to your library's OPAC? Or will records of the items in the repository be kept separately from the items in the OPAC?

What sort of persistent identifier issues arose during the course of the project? Which form of purl has your project chosen, and why?

Several ejournal projects have utilized Digital Object Identifier (DOI) links provided by publishers, but other projects have discovered issues related to providing a permanent link to an item housed in a digital repository which is actually a preserved copy of another item held by a publisher. How can digital repositories reconcile similar items with different identifiers, especially in increasingly networked repository environments?

Does your repository architecture provide any network capabilities with other OAIS archives? How will OAIS repositories approach the sharing of information resources? Please discuss ways your repository might be a part of a distributed environment. This would include a discussion of mirroring technologies such as rsync, LOCKSS and the Open Archives Initiative (OAI)...

The above discussion leads to a discussion of whether or not the archive would allow access to intelligent agents or bots? Will the repository allow this? This also might be a place to speak generally about the security of the repository...

If the repository architecture is still to be implemented, what will it take to get the architecture built in terms of dollars, peoplepower, expertise, and time?

Final Overview of Project:

Do you feel that your project has been successful to this point? In what ways was it successful? It what ways did it fall short of the goals that were set at the outset?

The Harvard ejournal archiving project identified the overarching need to develop standards for publishers and repositories in order to facilitate successful information exchange in the future. Work on the global digital format registry, the E-journal archive DTD, and the SIP specification are three areas where Harvard research is directly helping to create these standards. If your project has been working on developing standards, please discuss the progress on the pieces of "sharable Infrastructure" that you are working to develop. How important is the development of standards for each piece of the OAIS puzzle in ensuring its success?

Almost all of the digital repository projects have expressed concern that the funding requirements for the construction of repositories are so great that it is difficult for any single institution to make them a reality. Can you comment on the costs of such projects, as well as the potential sustainability issues?

Many of the repository projects expressed concern that there were no business models to support their attempts to determine the potential costs of creating a digital repository. Should work on business models for sustainable digital repositories be one of the most important research priorities?

The Harvard Report on the Planning Year Grant (from the Mellon ejournal archiving project) advocates largeness of scale as one way to make archiving of this sort economically feasible. A position which advocates for large-scale repositories brings into question the viability of self-archiving if not accompanied by supporting infrastructure. Discuss the advantages of scale in relation to your project.

Related to the above are conversations I've had with faculty members which suggest that they would find self-archived materials (even those self-archived under the auspices of parent institutions) to be inherently suspect because they have not undergone peer review outside of the individual institution. Is this a significant concern for institutions pushing to self-archive their own research materials?

Generally speaking, how do you view the current state of digital preservation initiatives? Is the research community close to finding solutions to some of the bigger problems, or are real solutions a long way off? Is enough funding generally available to tackle the issues that need to be tackled?

What questions did I not ask that you think I should ask?

Who else should I interview about these issues?

Glossary

Alphabetical List of Terms

Access-(1)The OAIS entity that contains the services and functions which make the archival information holdings and related services visible to consumers. [2]: (2) To make available (in accordance with all applicable restrictions) records, copies of records or information about records through activities such as reference services, providing reproductions, and producing exhibitions and public programs. [3]

Accession-To transfer physical and legal custody of documentary materials to an archival institution. [1]

Administration-(1) The services and functions that control the operation of the repository system. The services and functions include the capability to control submission, authentication requests for access, control access to digital objects, audit the access of digital objects, manage the repository system configuration, monitor the performance of the repository system, and provide support service activities related to the original collections and their reformatting, such as the compilation of statistics and usage reports. [4]: (2) The OAIS entity that contains the services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis. [2]

Administrative Metadata-Data that supports the unique identification, maintenance, and archiving of digital objects, as well as related functions of the organization managing the repository. Administrative metadata includes data identifying the owner and the provenance of a digital object, enabling use management of a digital object and supporting the migration of digital objects from one format to another for long-term preservation. [4]

Administrative Value-The value of records for the ongoing business of the agency of records creation or its successor in function. [1]

Appraisal-The process of determining the value and thus the disposition of records based upon their current administrative, legal, and fiscal use; their Evidential and Informational value; their Arrangement and condition; their Intrinsic Value; and their relationship to other records. [3]

Archival Information Package (AIP)-An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS. [2]

Archival Value-The value of documentary materials for continuing preservation in an archival institution. [1]

Arrangement-The archival process of organizing documentary materials in accordance with archival principles. [1]

Audit Trail-Information stored in the system log that provides the capability to discover an action or series of actions taken by the system, including actions initiated by either the system or by an individual interacting with the system. [3]

Authenticity-The property of a record that it is what it purports to be and has not been corrupted. [3]

Bit -A bit (short for binary digit) is an information unit used in computing and information theory. It is the smallest unit of storage currently used in these fields. A single bit is a 0 or a 1, or a true or a false, or any two mutually exclusive states. A byte is a collection of bits, originally variable in size but now usually eight bits. [8]

Bitstreams-The sequences of zeroes and ones that comprise data. [4]

Common Services-The supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, and directory services necessary to support the OAIS. [2]

Content-Generic term for data and metadata stored in the repository, individually or collectively described. [4]

Content Information-The set of information that is the original target of preservation. It is an Information Object comprised of its content data object and its Representation Information. [2]

Context-The organizational, functional, and operational circumstances in which a record is created and/or received and used. [3]

Custody-(1) Guardianship, or control, of records, including both physical possession (physical custody) and legal responsibility (legal custody), unless one or the other is specified. [3]: (2) The archival principle that to guarantee archival integrity, archival materials should either be retained by the creating organization or transferred directly to an archival institution. [1]

Data-A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen. [2]

Deed of Gift-A legal document accomplishing donation of documentary materials to an archival institution through transfer of title. [1]

Descriptive Metadata-Data that describes the digital object. [4]

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. [2]

Digital Object- A data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a handle (and, perhaps, other material). [9]

Dissemination Information Package (DIP) -the Dissemination version of an Archival Information Package (AIP). [2]

Digital Preservation-The managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents. [6]

Digital Repository-The facilities, personnel, processes, systems, and media used to store, manage, and provide access to digital objects. [4]

Emulation-The creation of an artificial environment within a new generation of technology that allows processes and data from an older generation of technology to exist and perform in their native format. [4]

Enduring Value-Stems from a document or record's intrinsic attributes, the contextual documentation that surrounds it, its relationship to other records and entities, and assurance of its authenticity and reliability. [11]

Essence-The bitstreams within a digital object that represent sound, texts, or still or moving images. [4]

Evidence-The passive ability of documents and objects and their associated contexts to provide insight into the processes, activities, and events that led to their creation for legal, historical, archaeological, and other purposes. [7]

Evidential (or Evidentiary) Value-The value of records or papers as documentation of the operations and activities of the records-creating organization, institution, or individual. [1]

Finding Aid-A description from any source that provides information about the contents and nature of documentary materials. [1]

Fixity-Authentication mechanisms and keys to ensure that the essence and metadata of a record have not been altered in an undocumented manner. [4]

Handle-A persistent name (Uniform Resource Identifier, or URN) in a form developed by the Corporation for Research Initiatives (CRN).

Hierarchical Description-The principal of archival description in which records are described in aggregates at various prescribed hierarchical levels. These levels range from the largest grouping (series) to the intermediate level (file unit) to the smallest (item). Descriptions of records at the series level are also linked to one of two types of archival control groups: a record group or a collection. [3]

Informational Value- The value of records or papers for information they contain on persons, places, subjects, and things other than the operation of the organization that created them or the activities of the individual or family that created them. [1]

Information Package-The Content Information and associated Preservation Description Information which is needed to aid in the preservation of the content information. The Information Package has associated packaging information used to delimit and identify the Content Information and Preservation Description Information. [2]

Ingest-The OAIS entity that contains the services and functions that accept Submission Information Packages from producers, prepares Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting Descriptive Information become established within the OAIS. [2]

Item-The lowest level of hierarchical description as defined by NARA, which describes the smallest intellectually indivisible archival unit. [3]

Levels of Control-The progressive grouping and describing of sets of records along a continuum from the largest and most general to the smallest and most specific. Thus the records of an agency can be successively both physically subdivided and intellectually described in terms of its constituent offices, activities, or functions; the files within each series; and the documents within each file. Each of these refinements is regarded as a different level of control. [12]

Life Cycle of Records-The concept that records pass through a continuum of identifiable phases from the point of their creation, through their active maintenance and use, to their final disposition by destruction or transfer to an archival institution or records center. [1]

Long Term-A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future. [2]

Long Term Preservation-The act of maintaining information, in a correct and independently understandable form, over the Long Term. [2]

Metadata-Data about other data. [2]

Migration-The act of moving electronic records and related data from one piece of media to another, usually in response to improving media technology, to avoid the inability to access records on media that is becoming obsolete, or to move records from media that is deterioration onto fresh media. [3]

Open Archival Information System (OAIS) -An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. [2]

Original Order-The archival principle that records should be maintained in the order in which they were placed by the organization, individual, or family that created them. [1]

Packaging Information-The information that is used to bind and identify the components of an Information Package. [2]

Permanent Value-The value of a record of sufficient historical or other value to warrant its continued preservation by the Federal Government beyond the time it is needed for administrative, legal or fiscal purposes. [3]: Contrast with Enduring Value.

Persistence-The attribute of essence that stands for long life expectancy; persistence is provided by appropriate management of the content. Content may persist when the underlying Bitstreams are maintained through time. In order to ensure the persistence of a representation of a library item, it may also be necessary to migrate or transform the component zeroes and ones. [4]

Persistent Format-A data type, which may be simple or complex, that is independent of specific hardware or software, such that an object in this data type can be transferred from a source platform to an arbitrary target platform with no significant alteration of essential attributes or behaviors. [3]

Personal Papers-A natural accumulation of documents created or accumulated by an individual or family belonging to him or her and subject to his or her disposition. Also referred to as Manuscripts. [1]

Preservation-Processes and operations involved in ensuring the technical and intellectual survival of authentic records through time. [3]

Primary Values-The values of records for the activities for which they were created or received. [1]

Processing-All steps taken in an archival repository to prepare documentary materials for access and reference use. [1]

Provenance-(1) The archival principle that records created or received by one recordskeeping unit should not be intermixed with those of any other. (2) Information on the chain of ownership and Custody of particular records. [1]

Record-A unit of recorded information of any type that is made or received in the course of activity and is kept because it provides evidence of the activity, is required by law or regulation, or contains valuable information. [3]

Reference-The ability to locate a digital object definitively and reliably over time among other digital objects. [5]

Reference Model-A framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist. [2]

Reliability-The authority and trustworthiness of records as evidence, the ability to stand for the facts they are about. A record is considered reliable when it can be treated as a fact in itself, as the entity of which it is Evidence. [10]

Representation Information-The information that maps a data object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of Bits (i.e., a data object) is mapped into a symbol. [2]

Respect des Fonds-The archival principle that records should be grouped according to the nature of the institution that accumulated them. [7]

Secondary values-The values of records to users other than the agency of record creation or its successors. [1]

Self-Describing-An entity whose data structure, format, or layout provides both definitions and values for the data or formats of the entity. A self-describing entity can be evaluated, with all its elements and formats understood, without the need of external references. [3]

Structural Metadata-Data that represents the relationships between components of complex multipart objects, e.g., the indication that this image represents "page one," this image "page two," and so on. Structural metadata supports the presentation and navigation of these objects. [4]

Submission Agreement-The agreement reached between an OAIS and the producer that specifies a data model for the data submission session. This data model identifies format/contents and the logical constructs used by the producer and how they are represented on each media delivery or in a telecommunication session. [2]

Trusted Digital Repository-A digital repository whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future. These repositories must meet eight criteria defined in the *Trusted Digital Repositories* report.

Uniform Resource Name (URN) -The URN is a persistent name, valid for the long term and independent of location, while still being globally unique. [4]

Sources for Glossary Terms:

[1] The *Appendix* and *Glossary* from Maygene F. Daniels and Timothy Walch, eds. *A Modern Archives Reader : Basic Readings on Archival Theory and Practice.* Washington, D.C. : National Archives and Records Service, U.S. General Services Administration. 1984. Their glossary draws on "A Basic Glossary for Archivists, Manuscript Curators, and Records Managers," compiled by Frank B. Evans, Donald F. Harrison, and Edwin A. Thompson, published in *The American Archivist* 37, July 1974, pages 415-433.
[2] Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS)." CCSDS 650.0-B-1. Blue Book. CCSDS: Washington, D.C. 2002. Available at http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf. Pages 1-7 through 1-13.
[3] *National Archives and Records Administration Request for Proposal (RFP*). 24 December 2003. NAMA-03-R-0018 for the Electronic Records Archives Amendment 0001. Washington, D.C.: National Archives and Records Administration. Available at http://www.archives.gov/electronic_records_archives/pdf/rfp.pdf. Found in *Appendix A: Glossary of Terms*, pages J2-A-1 through J2-A-13.
[4] *Library of Congress Digital Audio-Visual Repository System (DAVRS) System Requirements Document.* LC-DAVRS-02. Washington, D.C.: Library of Congress. 2000. Found in *Appendix A: Glossary*, pages 23-27.
[5] Garrett, John and Donald Waters, eds. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Mountain View, CA: Research Libraries Group. May 1996. Available at http://www.rlg.org/ArchTF/.
[6] *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report.* Mountain View, CA: Research Libraries Group. 2002. Available at http://www.rlg.org/longterm/repositories.pdf.
[7] Gilliland-Swetland, Anne J. "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment." Washington, D.C.: Council on Library and Information Resources. February 2000. Available at http://www.clir.org/pubs/reports/pub89/pub89.pdf.
[8] Wikipedia, the Free Encyclopedia: http://en.wikipedia.org/wiki/Main_Page.
[9] Kahn, Robert and Robert Wilensky. "A Framework for Distributed Digital Object Services." 1995. Available at http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.
[10] Duranti, Luciana. "Reliability and authenticity: the Concepts and their Implications." *Archivaria* no. 39 (Spring 1995).
[11] Tibbo, Helen R.. "On the Nature and Importance of Archiving in the Digital Age." *Advances in Computers.* v. 57, 2003: 1-67.
[12] Miller, Fredric M. *Arranging and Describing Archives and Manuscripts*. Chicago : Society of American Archivists. 1990.