**CITATION**
Ward, J.H. (2012). *Managing Data: the Emergence & Development of Digital Curation & Preservation Standards*. Unpublished manuscript, University of North Carolina at Chapel Hill.

## ABSTRACT

Archivists, librarians, computer scientists and other researchers and scientists have been concerned about the long-term survivability of data for decades. This data may be in the form of actual data sets, or data that represents and describes published works, art, video, audio, or other file formats. This literature review describes the emergence of digital curation and digital preservation standards in the context of managing data. Standards for digital curation and digital preservation augment the ability of data owners and users to ensure the survivability of their data, but these standards do not directly “cause” the long-term preservation of the data itself. The conclusion is that the survivability of data depends on the will and desire of the data owners and users, and the availability of financial resources to do so.

## TABLE OF CONTENTS

ABSTRACT .......... 2
INTRODUCTION .......... 5
WHY PRESERVE AND CURATE DATA? .......... 7
BASIC DEFINITIONS .......... 8
MOTIVATING FACTORS FOR THE DEVELOPMENT OF DIGITAL CURATION & DIGITAL PRESERVATION STANDARDS .......... 15
PERSISTENCE .......... 17
OVERVIEW OF THE OAIS REFERENCE MODEL AND THE AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES RECOMMENDED PRACTICE .......... 28
APPLICATIONS OF THE AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES RECOMMENDED PRACTICE AND THE OAIS REFERENCE MODEL .......... 38
OTHER TECHNICAL ISSUES .......... 43
GENERAL DIGITAL REPOSITORY MANAGEMENT .......... 46
FUNDING .......... 47
CURRENT STATUS AND FUTURE CHALLENGES/FURTHER WORK .......... 48
CONCLUSION .......... 50
REFERENCES .......... 52

## TABLE OF FIGURES

Figure 1 - the Digital Curation Centre Curation Lifecycle Model (Higgens, 2007). ......... 13
Figure 2 - Scientific and Technical Information Lost Over Time (Nelson, 2000)............. 20
Figure 3 - Core data grid capabilities and functions for implementing a persistent archive (Moore & Merzky, 2002)............................................................................... 24
Figure 4 - the OAIS Reference Model "Functional Model" (CCSDS, 2002)................... 30

## INTRODUCTION

Librarians and archivists have spent centuries wringing their hands and having multiple heated discussions about the best method for recording information for the purposes of transmitting it to succeeding generations. It could be, perhaps, that librarians and archivists of ancient Egypt were in agreement that clay tablets were the best form to transmit information, but one imagines that even within that framework there was much discussion with many agreements and disagreements. Which clay would last? Who was the best potter to fire it? How do we know that potter is actually selling us the quality of clay that he promised? "There must be standards!"…and so on and so forth. Regardless, those ancient librarians and archivists chose well, as 5,000 years later, those clay tablets have stood the test of time and are still readable now (Krasner-Khait, 2001).

Then "someone" determined that papyrus was better than clay as an information transmission form. After all, it was lighter, it couldn't break, and it was much easier to carry over long distances. The material required less storage space, as well, which would reduce overall costs. One can imagine the "old school" librarians and archivists with their clay fetish, snubbing the new papyrus advocates. However, the papyrus advocates eventually won, and the rolls of papyrus replaced clay tablets as the information medium of choice. Papyrus remained the primary information storage method of choice for around 3,000 years, until the development of the codex by the Romans in the first century A.D. (Zen College Life, 2011).

One can only imagine the consternation old school papyrus librarians and archivists faced with the invention of the codex. Should they change all of their holdings of clay tablets and papyrus rolls to codices? Should they leave this information in the old technologies and only store new information in the codex format? How many resources of time, money, and personnel would it take to migrate information from the old formats to the new? By 300 A.D., the codex was as popular as the papyrus scroll -- and the first and current format used for the Christian Bible. These debates, and one can be sure there were discussions, were not purely academic. There were then, as now, practical reasons to be concerned with the transmission of historical, cultural, political, and literary information to succeeding generations. By the time Gutenberg invented the movable type press in the 15th Century, the codex had evolved into the book, and another information revolution occurred. Books became more prevalent, and no doubt librarians and archivists of Western Europe, Asia and the Middle East felt an information deluge of their own as they figured out how to organize, lend, copy, store, and find these books as libraries and archives grew and evolved from the middle ages to the 20th Century (Zen College Life, 2011).

The mid-20th Century brought the computer, and then networked computers that share and store information as bits and bytes. The formats these bits are stored in evolve every few years, as do the software to run the formats, and the hardware that runs the software. Format changes now occur every few years, and make the 3,000 year reign of clay tablets as the information transmission form of choice unimaginable. Yet, one is certain that current librarians and archivists are solving the same problems their counterparts faced 5,000 years ago. How do you select, preserve, maintain, collect and archive information in order to make it available to succeeding generations? This is the essence of curation, whether digital or physical. However, the focus of this paper is to discuss the curation and preservation of binary data; therefore, curation methods as applied to physical artifacts are out of the scope of this discussion.

## WHY PRESERVE AND CURATE DATA?

There are many, many motivations for preserving data, regardless of the content. It would be challenging to cover every possible reason why some person or organization might want to curate and preserve their data. A few themes are common, though. In some instances, preservation is motivated by the human desire to preserve the current record (in a general sense) for future generations to access and use. Other motivations may be more base -- to help a particular company or organization comply with legal requirements or provide a source of revenue. In some cases, cultural heritage concerns may overlap with financial incentives, such as with digital movies. For example, executives at movie companies have a huge financial incentive to ensure that their libraries are accessible in the future as formats change, so that they may sell and re-sell their titles for public consumption (Science and Technology Council, 2007). These films also represent the cultural heritage of humanity, whether the film in question is "Harold and Kumar Take Guantanamo Bay" or "Citizen Kane". In other organizations such as the National Archives, federal legal requirements overlap with a professional desire and charge to preserve the United States' materials "for the life of the republic" (Thibodeau, 2007). Individuals' health records must be available for the life of the person. Most of us would like our photographs to be accessible by our descendants and relatives, and not lost in a hard drive or a hard drive crash. These are but a few examples of "what" and "why" data are deemed preservation-worthy.

## BASIC DEFINITIONS

*"Archive", "digital archive", "data", "information", "knowledge", "wisdom", "digital preservation", "digital curation", "reliable", "authentic", "integrity", and "trustworthy".*

Tibbo (2003) writes that computer scientists tend to use "archive" simply as a term to describe the storage and backup of digital data in an offline electronic environment, while archivists see the process of archiving data as part of a complex process that encompasses the entire lifecycle of a digital object (Waters & Garrett, 1996; Higgens, 2007). One may also see the difference between "archive" in the computer science sense as simply storing data, whereas an "archive" per an archivist is an entire information system lifecycle that encompasses data, information, knowledge, and, perhaps, wisdom that will be made accessible for the indefinite long-term.

As well, practitioners who work with digital libraries and digital archives often use "digital library" to mean a "digital archive", and vice versa. What then, is a digital archive?

Waters and Garrett (1996) defined

> "digital archives strictly in functional terms as repositories of digital information that are collectively responsible for ensuring, through the exercise of various migration strategies, the integrity and long-term accessibility of the nation's social, economic, cultural and intellectual heritage instantiated in digital form. Digital archives are distinct from digital libraries in the sense that digital libraries are repositories that collect and provide access

> to digital information, but may or may not provide for the long-term storage and access of that information. Digital libraries thus may or may not be, in functional terms, digital archives and, in fact, much of the recent work on digital libraries is notably silent on the archival issues of ensuring long-term storage and access….Conversely, digital archives necessarily embrace digital library functions to the extent that they must select, obtain, store, and provide access to digital information. Many of the functional requirements for digital archives defined in this report thus overlap those for digital libraries."

The Society of American Archivists (1999) defines the core curation functions of any archive as appraisal, accession, arrangement, description, preservation, access and use. The basic archival principles remain the same whether an archive contains physical artifacts or data (Hedstrom, 1995). How an archivist applies these concepts may vary depending on the digital objects or physical artifacts to be preserved. Within the limitations of digital data, however, most applications of a data archive as of this writing use the Open Archival Information System (OAIS) (Consultative Committee for Space Data Systems, 2002) as a reference model. This model will be discussed briefly in a later section. However, the Consultative Committee for Space Data Systems (2002) notes that an "OAIS Archive" is distinguished from other uses of the term "archive" because it consists of an "organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community". The archive must meet the set of responsibilities outlined in the OAIS Reference Model to be considered an "OAIS archive" (Consultative Committee for Space Data Systems, 2002). Otherwise, it is merely an "archive".

Data are "any information in digital form" (Higgens, 2007) that "correspond to the bits (zeroes and ones) that comprise a digital entity" (Moore, 2002). Data include both simple and complex objects, as well as structured collections. A simple object may be a text file or image; a complex file may comprise an entire web site; and a database is an example of a structured collection (Higgens, 2007). Furthermore, Galloway (2004) notes that to be digital the objects must "require a computer to support their existence and display".

Moore (2002) writes from a computer science perspective that information "corresponds to any tags associated with bits", while Buckand (1991) defines information via the lens of Information Science. He describes "information-as-process", "information-as-knowledge", and, "information-as-thing". According to Buckland, "information-as-process" is the act of informing, while "information-as-knowledge" is the actual knowledge communicated during "information-as-process". He defines "information-as-thing" by objects such as text and data, for example, because they impart and communicate knowledge; and notes that knowledge may be contained in text, etc. that describes these information objects. Ackoff (1989) takes a management science approach and posits that information is contained in answers to questions posed with "who", "what" "where", and "when".

Knowledge "corresponds to any relationship that is defined between information attributes" (Moore, 2002); it is the application of data and information. Knowledge refines information and makes "possible the transformation of information into instructions" by answering the "how" questions (Ackoff, 1989). Wisdom is at the pinnacle of Ackoff's hierarchy as an ideal state that evaluates the long-term consequences of an act. One might argue that repositories with audit mechanisms to ensure "authenticity" and "trust" apply wisdom in the form of policies to curate data, information, and knowledge "as things".

The phrases "digital curation" and "digital preservation" are often used interchangeably, but they have slightly different meanings. The term "digital preservation" refers to a "series of managed activities necessary to ensure continued access to digital materials for as long as necessary" (Digital Preservation, 2009). Members of the Digital Preservation Coalition made this definition deliberately broad in order to refer "to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change" (Digital Preservation, 2009). As part of that definition, these members framed digital preservation as short-term, medium-term, and long-term. They defined "short-term" as access to the materials for the foreseeable future or for a defined period of time; "medium-term" as providing access for the near-term but not indefinitely; and, "long-term" as providing continued access to the materials for the indefinite future (e.g., as long as possible). Hedstrom (1995) writes that

> "preservation of an electronic record entails retaining its content; maintaining the ability to reproduce its structure; and providing linkages between an archival document and related records, its creator and recipient, the function or activity that it derived from, and its place in a larger body of documentary evidence."

Researchers and practitioners at the Digital Curation Centre (DCC) have defined digital curation as involving "maintaining, preserving and adding value to digital research data throughout its lifecycle" (Digital Curation Centre, 2010). An archivist, librarian or other data manager begins curation at the time the collection is assembled or acquired. He or she actively manages the collection in order to "mitigate the risk of digital obsolescence" and "to reduce threats to [the data's] long-term research value" (Digital

Curation, 2010). According to DCC researchers and practitioners, curation serves two other primary purposes that include providing a means to share data and reducing duplication of effort in data creation.

Higgens (2007) conceptualized an ideal model of digital curation as a lifecycle with three primary areas: full lifecycle actions, sequential actions, and occasional actions. These actions may be applied across the entire digital lifecycle or sequentially through it (Higgens, 2007). She defines "full lifecycle actions" as encompassing preservation planning; description and representation information; and, curation and preservation. Higgens models sequential actions as: conceptualization; creation or reception; access, use, and re-use; appraisal and selection; ingestion; storage; preservation action; and, transformation.

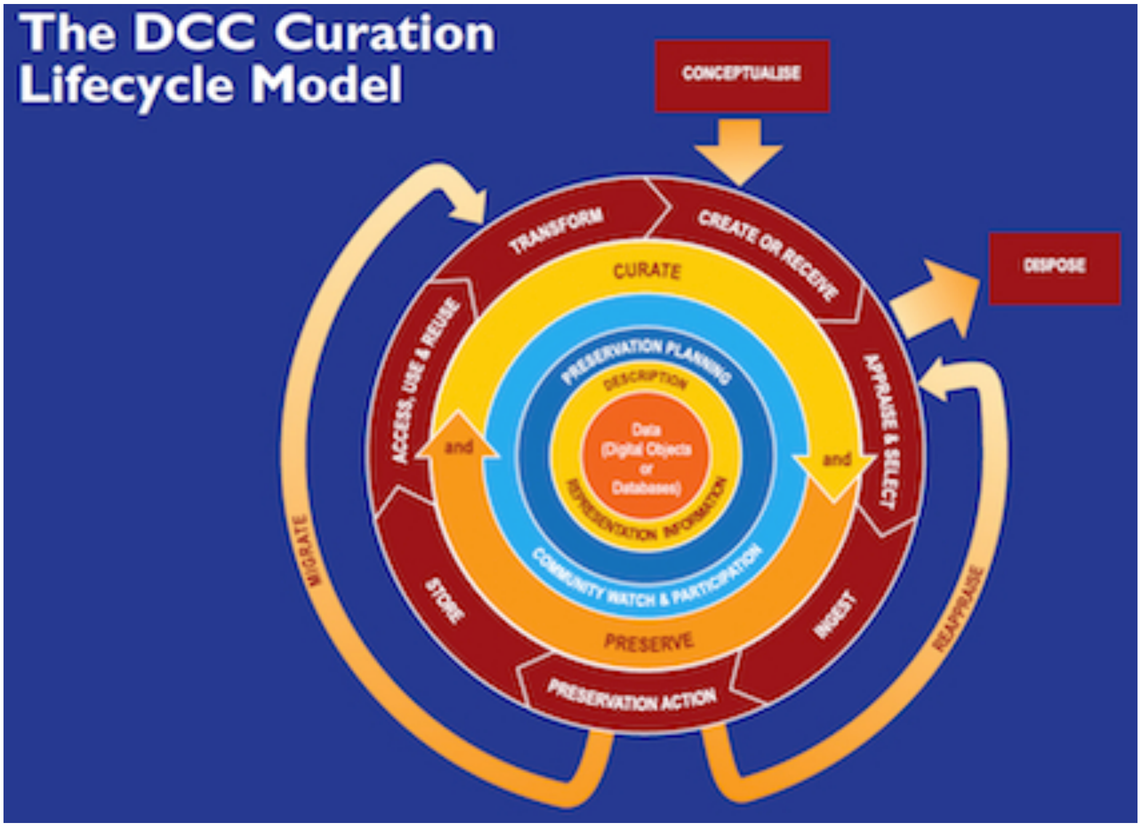


Figure 1 - the Digital Curation Centre Curation Lifecycle Model (Higgens, 2007).

The significance of the model is that provides a visual tool and summary from which a repository manager may plan the curation tasks appropriate for the collection and the repository at any stage in the curation lifecycle.

Duranti (1995) defined the terms "reliability" and "authenticity" based on diplomatic concepts. A record is "reliable" when the degree of completeness of the form and the degree of control of the procedure of creation meet the requirements of the socio-juridical system in which it is created. A reliable record is a "fact in itself, that is, as the entity of which it is evidence" (Duranti, 1995). If a document is what it claims to be, then the document is considered authentic. However, just because a document is authentic does not make it reliable. If a record is authentic, then it "does not result from any manipulation, substitution, or falsification occurring after the completion of its procedure of creation" (Duranti, 1995). Reliability takes precedence over authenticity.

The way to guarantee both reliability and authenticity is to have a standard for record completeness along with a controlled procedure for creation as well as a procedure to control the transmission and storage of the records. For example, a birth certificate will be considered reliable and authentic if all fields required by law have entries, the person providing the information has the authority to do so (i.e., is the attending physician or midwife) from a knowledgeable source (i.e., one or both parents as well as their own attendance at the birth), the authorized person enters the information provided correctly, the parents provide the correct information to begin with, and the birth certificate is stored in a government repository with access controls to the repository records. If a parent or physician provides false information on the birth certificate and the government stores it, then subsequent copies obtained of the birth certificate may be authentic, but they will not be reliable.

In order to provide reliable, authentic records in a digital environment, the keepers of the data objects must be able to maintain the objects' integrity and provide evidence that that repository itself is trustworthy. The primary evidence of an objects' integrity relate to its content, fixity, reference, provenance, and context (Waters & Garrett, 1996). Integrity builds upon, and to some degree, is concerned with authenticity, but it is not security (Lynch, 1994). Some examples of integrity violations include bit flipping, data corruption, disk errors, and malicious intrusions (Sivathanu, Wright, and Zadok, 2005).

At a minimum, for a repository to be trustworthy, it must begin with "'a mission to provide reliable, long-term access to managed digital resources to its Designated Community, now and into the future'" (Consultative Committee for Space Data Systems, 2011). Both Waters & Garrett (1996) and the Consultative Committee for Space Data Systems (2011) prefer that repository managers conduct transparent audits of the system itself in order to assure "trustworthiness" to both internal and external stakeholders.

## MOTIVATING FACTORS FOR THE DEVELOPMENT OF DIGITAL CURATION & DIGITAL PRESERVATION STANDARDS

The movement to set standards for preservation and curation developed to provide order to chaos, and provide the information necessary so that individuals and organizations may make informed decisions about which data objects are reliable and authentic, and which repositories are trustworthy and mindful of data object integrity. That is, practitioners need to be able to determine if the people running a repository are actually doing so in a way that will preserve the objects for the specified time required in such a way that those objects can be found. More importantly, practitioners and users also must be certain that the objects preserved are both authentic and reliable. One way to ensure the reliability, authenticity, integrity, and trustworthiness of data objects and the repositories that house them are for the stakeholders to come together and agree on the procedures and definitions for those, and in the process, create standards for digital curation and digital preservation.

Previously, different industries worked within their domain to develop standards for preservation and curation. Book publishers worked within book publishing; filmmakers within filmmaking, and so on and so forth (Science and Technology Council, 2007). The mass use of digital data has created the need for broad standards that cross all industries. This is not a situation where knowledge about how to preserve one kind of format tends to exclude knowledge of how to preserve another kind of format, e.g., paper vs. film. A digital file is a digital file, whether it resides in a repository at the Library of Congress or in a graphic designer's personal laptop. All industries are facing similar problems; a short list of these problems include format obsolescence, physical media changes, hardware and software migrations, personnel costs, as well as the costs of storing all of this data for perpetuity and making it accessible.

The latter -- cost -- ranks among one of the highest concerns. For example, the cost of storing a 4k digital master of a movie is 1,100 times higher than storing the same master as film (Science and Technology Council, 2007). A collection may be deemed worthy of saving into perpetuity by a consensus of experts, but without any resources to make that happen, the most one can hope for is that the machine the data is stored on will be turned off and put in a temperature-controlled closet until and if "someone" finds it and migrates the data to a new resource. (This preservation method assumes the data can be migrated and that there has not been any physical deterioration of the machine or disks, etc., during the time it was in storage.)

What is the best way to reduce long-term preservation costs? According to the members of the Science and Technology Council of the American Academy of Motion

Picture Arts and Sciences (2007), the best way to reduce costs is to collaborate within and across industries and domains to develop and use standards, leveraging organizations such as the National Digital Information Infrastructure & Preservation Program (NDIIPP) for this purpose. The word "standard" includes, but is not limited to, file formats, filenames, metadata, metadata registries, distribution and archiving. Gallloway (2004) also concluded that the costs of preserving digital materials are exacerbated by the proliferation of proprietary formats, and that the format problem must be solved in order to limit cost.

## PERSISTENCE

As stated earlier, digital curation and preservation standards grew out of established practices for the preservation of the human record, whether the purpose is research, legal requirements, cultural heritage, etc. One idea behind the development of standards, best practices, reference models, audit criteria, and a lifecycle model, etc., is to create a body of knowledge such that any person charged with preserving and curating a digital collection may readily find the information needed to accomplish their task.

Waters & Garrett (1996) were part of the Task Force on Archiving of Digital Information that examined the "state of the state" of digital preservation in the mid-1990s. Many of the task force's recommendations contributed to the development of the final versions of the OAIS (Consultative Committee for Space Data Systems, 2002) and the standards for the Audit and Certification of Trustworthy Digital Repositories (Consultative Committee for Space Data Systems, 2011). Other recommendations from the 1996 task force include: creators, providers and owners of digital information are responsible for the preservation of the information; deep digital infrastructure must be developed to support a distributed preservation system; and, trustworthy, certified archives must be prepared and able to aggressively rescue data from repositories that are failing (Waters & Garrett, 1996).

While many large datasets have been preserved for decades without any formal standards for preservation and curation, it helps to have best practices with which to build a preservation program. For example, the Inter-University Consortium for Political and Social Research (ICPSR) has been migrating data since at least the early 1960s with few formal preservation criteria or curation standards to reference (Galloway, 2004). ICPSR personnel, partners, and users were committed to the longevity of the data, so it has been migrated repeatedly. Over the past few years, ICPSR has formalized their repository design to comply fully with the OAIS reference model, for example, because data managers believe this will further ensure the long-term availability of the social science data in the repository and lead to a "federated system of social science repositories" (Vardigan & Whiteman, 2007).

This year, Paul Ginsparg, physicists, mathematicians, computer scientists, and other scientists celebrated the 20th anniversary of arXiv, a pre-print archive (Ginsparg, 2011). Ginsparg began arXiv as an electronic bulletin board to continue physicists' tradition of sharing research via mail and email. The bulletin board grew into a digital repository, and has survived a variety of funding sources, media, hardware, and software changes. The creators of arXiv and affiliated researchers have used it as a test bed from which to create a variety of standards that have aided in repository architecture design and interoperability such as the Dienst Protocol and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Davis & Lagoze, 2000; Lagoze & Van de Sompel, 2001). Thus, practitioners had not yet created preservation and curation standards for repositories at the time of arXiv's birth, yet it has survived for 20 years because the community that uses it wants to keep using it.

Although there are many technical problems associated with digital preservation that have yet to be solved, including the rapid obsolescence of software and hardware due to technology cycles (Thibodeau, 2002; Rothenberg, 1999), the primary problems associated with digital preservation and the curation of data are not technical, they are societal. Galloway (2004) notes that whether or not data are preserved has more to do with whether or not a given community chooses to preserve its own record; intellectual and social capital are the issue. Although we are in the midst of a data deluge that is not going to grow smaller any time soon, if ever, there are adequate systems and designs to support it. There must be an institutional commitment to support the preservation of a particular set of data, and this commitment must include an expenditure of resources, not just of will or desire for digital preservation (Consultative Committee for Space Data Systems, 2002). Galloway (2004) lists organizations that have consistently migrated data due to institutional will and personnel commitment, and these include science (for data sets), data warehouses, publishers, including authors (text files), and government agencies (e.g., the National Archives, the Library of Congress, and other federal and state agencies).

Plenty of data has been lost over the years, as well, by those same organizations. Rothenberg (1999) listed several cases of possible loss by U.S. government agencies; one of the more famous examples is the census data for 1960 (although Waters and Garrett (1996) note that the data loss was not as extensive as some think). He points out that computer scientists are notorious for accepting data loss as part of the price one pays to move to the next generation of hardware and software. He also writes that in 1990, a Congressional report "cited a number of cases of significant digital records that has already been lost or were in serious jeopardy of being lost". To put this in perspective on a smaller scale, Nelson (2000) wrote that in a typical project at NASA c. 2000, the published research paper went to a library, the software to an FTP site, raw data was thrown away, and images were stored in a filing cabinet.

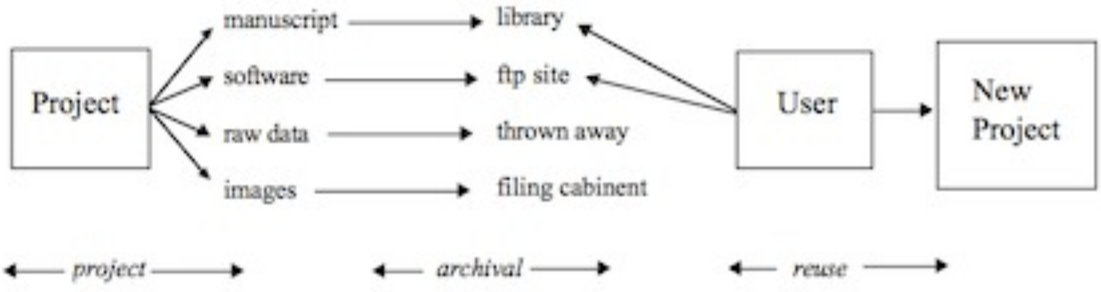


Figure 2 - Scientific and Technical Information Lost Over Time (Nelson, 2000).

In theory, all digital data could be preserved, but then the question becomes, "Should it be preserved?" If not, how do you cull that much data? Maybe it is better to keep it all? It takes personnel time to cull data, but it costs to store data that should otherwise be deleted, too.

The idea of how permanent or impermanent an archive's collections should be is not a new one. O'Toole (1989) wrote that archivists had evolved their attitudes towards the "permanence" of artifacts and had begun to view permanence as an "unrealistic and unattainable" ideal. This is echoed in the digital realm. The members of the InterPARES (2001) project determined that it is acceptable to preserve a version of a record, so long as the integrity of the information is maintained. In other words, if a file format must be migrated from one form to another (.doc to .pdf, for example) in order to preserve it, an archivist does not have to preserve the original bits for the information itself to be considered authentic. Thibodeau (2002) also noted that it is more important to preserve the essential characteristics of an object -- its look, feel, and content, for example -- than it is to preserve the digital encoding of the object per se. All preservationists do not share this view, however. As late as 1999, Rothenberg (1999) expected documents to be preserved in their original bit form.

The members of the Science and Technology Council (2007) reached a conclusion similar to Thibodeau and the InterPARES members regarding film masters versus digital movie masters. The practice for the past 100 years or so has been to "save everything" when archiving a film. Thus, a director may go back 20 or 30 years later and create a new version of a movie, or film buffs with access to the film archive may study other aspects of the movie itself. The council members concluded that "save everything" is not feasible with digital movies, both due to the number and size of a digital movie, plus the cost of storing that much data over time. The digital movies will have to be migrated from the original file format, software, and hardware, to be compatible with new file formats, software, and hardware. This new version will supersede the old version of a movie, thus changing the idea of what is the actual canonical copy of a film. Therefore, the idea that the objects in a digital collection are ephemeral -- both in terms of whether or not data will be kept in the first place, and that the canonical digital version itself will evolve over time -- is an idea that has gained ground as digital curation and preservation have developed over the past decades

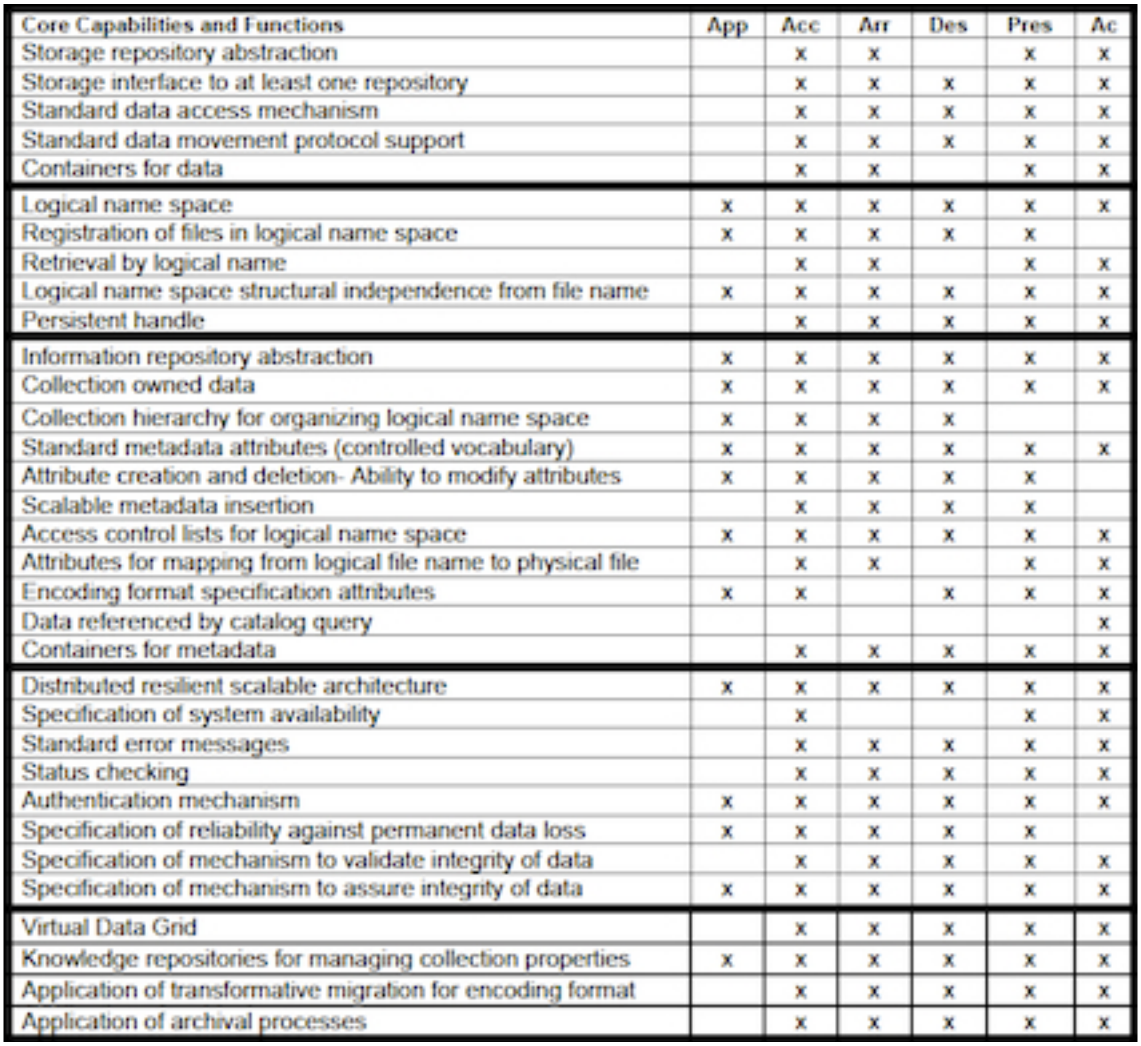
However, in spite of the idea that data are ephemeral either in terms of their lifespan or bits and bytes, another notion developed: that of "persistence". An archivist or computer scientist may not want to keep an object long, or he or she may wish to migrate the format, but he or she wants to be able to find that data and do what is needed to the object, whether that means deleting it, migrating it, or some other task.

One of the first tasks upon ingesting an object into a repository is to assign it a unique identifier that is not shared by any other object in the archive, and, preferably, by any object in any archive. A full discussion of unique identifiers is beyond the scope of this paper, much less a discussion of the pros and cons of the various identifiers available to use with data. Some unique identifiers are one-of-a-kind to the archive or archive owner only. Some are part of a larger standard, such as Digital Object Identifiers (DOI), which are persistent names linked to redirection (Paskin, 2003). Some identifiers work only with URIs and can only be used via the World Wide Web (WWW), such as ARK (Archival Resources Key) (Kunze, 2003).

Most identifiers used with digital data may be used as URLs/URNs (Universal Resource Locator/Universal Resource Name). These are web-based, and run over the

Internet. URLs are equivalent to a person's address (e.g., http://sils.unc.edu/), and URNs are the equivalent of a person's name, but the latter may be combined with existing non-Web identifiers to create a one-off, web-based identifier such as, "urn:isbn:n-nn-nnnnnn-n" (URI Planning Interest Group, 2001). Once a unique identifier is assigned, it is considered a best practice never to change that identifier, resource name, or resource locator (Berners-Lee, 1998). If it is necessary to do so for administrative or policy reasons, then within the system itself a "redirect" should be in place, so that the old location identifier points the system or user to the new location of the data.

As part of establishing persistent identifiers and locators for networked-based identifiers, researchers began to identify the features of a persistent (digital) archive, a persistent collection, and a persistent object. Moore & Merzky (2002) developed concepts for a persistent archive. They combined the functionality of a data grid with traditional archival processes (e.g., appraisal, accession, arrangement, and description) to create a matrix of core capabilities and functions.



| Core Capabilities and Functions | App | Acc | Arr | Des | Pres | Ac |
|---|---|---|---|---|---|---|
| Storage repository abstraction |  | x | x |  | x | x |
| Storage interface to at least one repository |  | x | x | x | x | x |
| Standard data access mechanism |  | x | x | x | x | x |
| Standard data movement protocol support |  | x | x | x | x | x |
| Containers for data |  | x | x |  | x | x |
| Logical name space | x | x | x | x | x | x |
| Registration of files in logical name space | x | x | x | x | x |  |
| Retrieval by logical name |  | x | x |  | x | x |
| Logical name space structural independence from file name | x | x | x | x | x | x |
| Persistent handle |  | x | x | x | x | x |
| Information repository abstraction | x | x | x | x | x | x |
| Collection owned data | x | x | x | x | x | x |
| Collection hierarchy for organizing logical name space | x | x | x | x |  |  |
| Standard metadata attributes (controlled vocabulary) | x | x | x | x | x | x |
| Attribute creation and deletion- Ability to modify attributes | x | x | x | x | x |  |
| Scalable metadata insertion |  | x | x | x | x |  |
| Access control lists for logical name space | x | x | x | x | x | x |
| Attributes for mapping from logical file name to physical file |  | x | x |  | x | x |
| Encoding format specification attributes | x | x |  | x | x | x |
| Data referenced by catalog query |  |  |  |  |  | x |
| Containers for metadata |  | x | x | x | x | x |
| Distributed resilient scalable architecture | x | x | x | x | x | x |
| Specification of system availability |  | x |  |  | x | x |
| Standard error messages |  | x | x | x | x | x |
| Status checking |  | x | x | x | x | x |
| Authentication mechanism | x | x | x | x | x | x |
| Specification of reliability against permanent data loss | x | x | x | x | x |  |
| Specification of mechanism to validate integrity of data |  | x | x | x | x | x |
| Specification of mechanism to assure integrity of data | x | x | x | x | x | x |
| Virtual Data Grid |  | x | x | x | x | x |
| Knowledge repositories for managing collection properties | x | x | x | x | x | x |
| Application of transformative migration for encoding format |  | x | x | x | x | x |
| Application of archival processes |  | x | x | x | x | x |

Figure 3 - Core data grid capabilities and functions for implementing a persistent archive (Moore & Merzky, 2002).

The authors proposed that this set of core capabilities would minimize the human labor involved in "implementing, managing, and evolving a persistent archive". More importantly, they noted that these capabilities already exist in (then) current implementations of data grids.

Moore (2005) evolved these ideas to include the concept of a "persistent collection". He defines a persistent collection as a "combination of digital libraries for the publication of digital entities, data grids for the sharing of digital entities, and persistent archives for the preservation of digital entities". Moore concluded that while persistent collections are built on top of data grids, and data grids have been used successfully for data sharing, publication, and preservation, in order to use data grids for persistent collections, additional capabilities "to simplify the integration of new services and support the federation of independent data grid federations" must be added.

Brody (2000) and Carr (1999) "mined" the life of an ePrint archive and discovered that authors still made corrections to the papers and metadata after the respective author or authors had submitted them to the University of Southampton ePrint archive. (Neither Brody nor Carr provided an average end date as to when authors stopped committing changes either to the paper or the metadata.) Thus, even Thibodeau's "essential characteristics" are subject to change, although the repository's owners could change this characteristic be creating a policy that allowed or prohibited changes post-publication in the repository.

Another aspect of object persistence is whether or not the Web site that contains the object or data currently exists (as opposed to available but not accessible). Koehler (1999) examined the persistence of Web pages, Web sites, and server-level domains beginning in 1996. He reported that after 6 months, 20.5% of Web pages and 12.2% of Web sites monitored for the study failed to respond. After 12 months, those figures changed to 31.8% and 17.7% respectively. He inferred from this that the half-life of a Web page is about 1.6 years, and a Web site, 2.9 years. Koehler determined 3 kinds of Web persistence: permanence (it is not going anywhere); intermittent (sometimes it is there, sometimes it is not); and, disappearance (it is gone forever). He discovered that

99% of Web sites had changed after 12 months. Koehler (1999) concluded that if the World Wide Web is the equivalent of H.G. Wells' (1938) "world brain", then two things may be said of it: the world brain has a short memory, and when it does remember, it changes its mind a lot -- how much and where depends.

Koehler (2004) revisited his study five years later. He reports that static collections -- similar to the ePrints archive mentioned earlier in this paper -- tend to stabilize after they have "aged". As part of this paper, he reviewed the growing body of literature related to persistence -- also referred to as "linkrot" -- and found that the stability of collection-oriented Web sites (e.g., legal, academic, citation-based) varies based on the domain specialty. Nelson and Allen (2002) examined 1,000 digital objects in a variety of digital libraries over the course of a year. They discovered that 3% of all objects were no longer available after 12 months, but the resource half-life is about 2.5 years. Koehler writes that for other resource types, such as scholarly article citations, legal citations, biological science education resources, computer science citations, and random Web pages, the half-life of the resources ranges between 1.4 years to 4.6 years. While some of the URLs in both of Koehler's studies stabilized for two years after losing two-thirds of the URLs in the first 4 years of the study, his overall conclusion was that the Web provides no guarantee of longevity for data, collections, or repositories.

McCown, Chan, Nelson, & Bollen (2005) revisited the Nelson and Allen (2002) study of D-Lib Magazine Web persistence and expanded upon it by examining outlinks -- the URLs cited in D-Lib Magazine articles. They extracted 4387 unique URLS referenced during July 1995 to August 2004 in 453 articles. They discovered that approximately 30% of URLs failed to resolve, although only 16% of the content registered indicated more than a 1 KB change during this same testing period. The researchers concluded that the half-life of a URL referenced in a D-Lib Magazine article is around 10 years. To state the obvious, even scholarly articles referenced in a respected journal in the Information and Library Science field -- where linkrot is a known problem -- cannot maintain stable references.

These studies above represent but a small proportion of the literature documenting the ephemeral nature of data ("digital objects"), Web sites ("archives"), and Web pages. By the late 1990s to early 2000s, it had became apparent in all fields that in order to rely on digital resources, some objects need to be static, the repository that contains the objects needs to remain accessible, there needs to be audit mechanisms to prove that the objects in the repository are what they say they are, and that the repository is capable of persisting over time even as the content is migrated to newer software and hardware. In other words, "someone" needed to develop a standard model for archiving objects for some period, either short- or long-term. As well, "someone" needed to create audit mechanisms to determine that a repository is "trustworthy" and that the repository's contents are "authentic" and "reliable" and have maintained their "integrity". "Someone" had been doing just that: the CCSDS finalized the "Reference Model for an Open Archival System" (OAIS) as a standard in 2002. The CCSDS released the "Audit and Certification of Trustworthy Digital Repositories" as a Recommended Practice (Magenta Book) in September 2011.

## OVERVIEW OF THE OAIS REFERENCE MODEL AND THE AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES RECOMMENDED PRACTICE

The Consultative Committee for Space Data Systems (CCSDS) convened an international workshop in 1995 with the purpose of advancing a proposal "to develop a reference model for an open archival information system" (Lavoie, 2004). The CCSDS had determined previously that there was no widely accepted model or framework that could serve as a standard for the long-term storage of space mission digital data. The members of the CCSDS recognized that fundamental questions related to digital preservation cut across all domains; therefore, the development scope of the model included stakeholders from a variety of domains, including government, private industry, and academia (Lee, 2010). The committee determined that the purpose of creating a reference model was to "address fundamental questions regarding the long-term preservation of digital material" (Lavoie, 2004). This model would define an archival system and outline the essential conditions a repository owner must meet in order to be considered a preservation archive.

The CCSDS defines an OAIS as "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. It meets a set of such responsibilities as defined in this document, and this allows an OAIS archive to be distinguished from other uses of the term 'archive'" (CCSDS, 2002). The word, "Open" is used to note that the CCSDS (2002) developed the recommendation in open forums, and will continue to do so for any future iterations of the model. The use of the word, "Open" does not imply that access to the repository itself must be unrestricted, in order to meet the requirements of the OAIS model (Lee, 2010).

The committee described four categories of archives: independent, cooperating, federated, and shared resources. The owners of an independent archive do not interact with any other archive owners with regards to technical or management issues. The possessor of a cooperating archive does not have a "common" finding aid with other archive possessors, but otherwise shares common producers, submission standards, and dissemination standards. The owners of a federated archive serve both a global and local Designated Community with interests in these related archives, and these owners provide access to their holdings to the Designated Community via one or more shared finding aids. The holders of archives with shared resources have agreed to "share resources" with each other, generally to reduce cost. This type of arrangement requires the use of standards internal to the archive, such as for ingest and access, that do not "alter the user community's view of the archive" (CCSDS, 2002; Lee, 2010).

The CCSDS divided the reference model into two "sub-models" - a Functional Model and an Information Model. Simply put, the Functional Model defines what an archive must do, and the Information Model defines what a repository must have in its collections (Lee, 2010). The former describes seven main functional entities, and however they interface with each other. These interfaces are: Common Services, Preservation Planning, Data Management, Ingest, Administration, Access, and Archival Services.

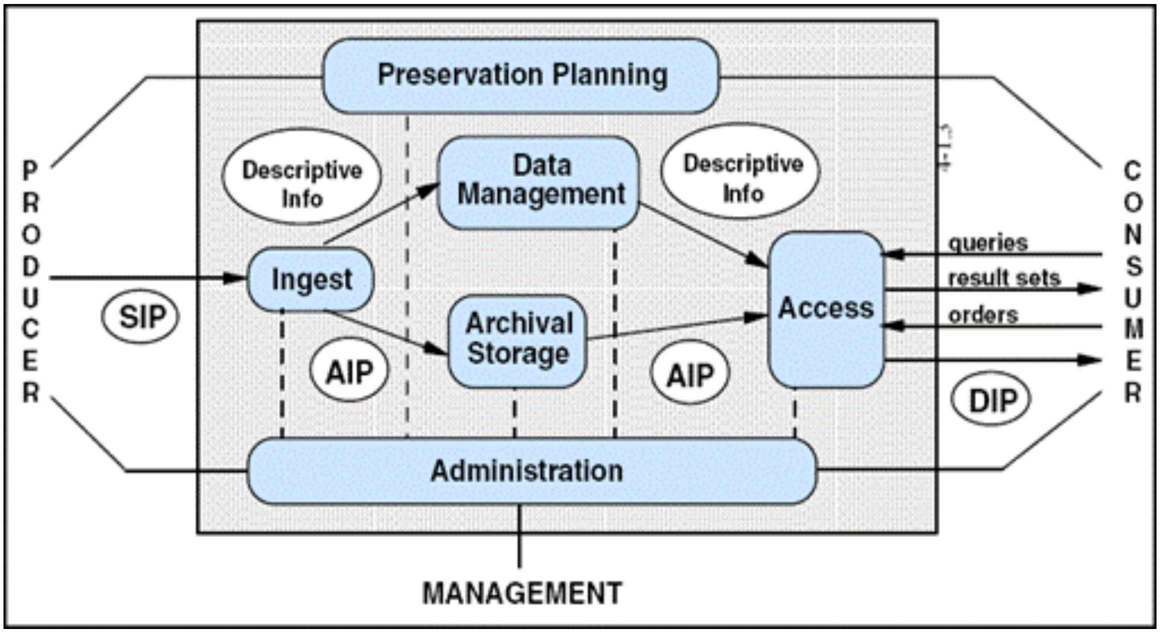


Figure 4 - the OAIS Reference Model "Functional Model" (CCSDS, 2002).

The Information Model describes and defines the information beyond the content. The members of the CCSDS included this section because the long-term preservation of digital material will require more than simply the content itself. A few examples of the information described and defined within the Information Model include: representation, fixity, provenance, content, and preservation description.

In summary, if a repository is an OAIS-type archive, then the archive managers will implement each area of the Functional Model in order to preserve information as an information package via the Information Model, for a Designated Community (Lavoie, 2004). The CCSDS designed the OAIS to be a reference model - it is NOT an implementation. The committee members deliberately left it up to an archive's owners to determine the technical details of the archival system. Egger (2006) writes that this is a disappointing aspect of the reference model, because it mixes technical and management functionality, rather than keeping them separate per standard software engineering practices. Vardigan and Whiteman (2007) successfully applied the OAIS reference model to the Inter-university Consortium for Political and Social Research (ICPSR) social science repository. The managers of the Online Computer Library Center (OCLC) Digital Archive based their service on the OAIS reference model while drawing data and metadata from a "wide array of OCLC organizational units" (Lavoie, 2004).

Another application of the conceptual work of the CCSDS (2002) with the OAIS reference model, and Waters and Garrett's (1996) work with the Task Force on Archiving of Digital Information, is the CCSDS' development of a "recommended practice" for the "audit and certification of trustworthy digital repositories" (CCSDS, 2011). This work is also based on the development of the requirement for a repository to be "reliable", "authentic", have "integrity", and, be "trustworthy", as defined in a previous section.

Lavoie (2004) writes that OCLC and the Research Libraries Group (RLG) sponsored an initiative in March 2000 to address the "attributes of a trusted digital repository". The working group's charge was "to reach consensus on the characteristics and responsibilities of trusted digital repositories for large-scale, heterogeneous collections held by cultural organizations" (Research Libraries Group, 2002). The purpose of determining these characteristics is to ensure that an OAIS Designated Community will be able to audit a repository and determine whether or not the repository owners have designed it, and are managing it, in such a way that the repository will actually preserve the Designated Community's data for the indefinite long-term and make it accessible. The RLG/OCLC working group issued their report in 2002. Among the recommendations, the working group specified that a process needed to be developed to certify a digital repository (Research Libraries Group, 2002). Waters and Garrett (1996) had also made this recommendation.

What is a "trusted digital repository"? The working group of the Research Libraries Group (2002) defined it as a repository with "a mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future". The NESTOR Working Group on Trusted Repository — Certification (2006) determined that the entire system must be looked at in order to determine whether or not a Designated Community should trust that it will last for the indefinite long-term. This includes its governance; procedures and policies; financial sustainability and fitness; organizational management, including employees; legal liabilities, contracts and licenses under which it operates; plus the trustworthiness of any organization or person who might inherit the data (NESTOR Working Group on Trusted Repository — Certification, 2006; Online Computer Library Center, Inc. & Center for Research Libraries, 2007).

A repository manager must also assess internal and external risks to the repository. Among many of the potential risks to a repository's long-term availability, Rosenthal, et al (2005) include internal and external attacks; natural disasters; hardware, software and media obsolescence; hardware, software, media, network services, organizational, and economic failure; as well as simple communication errors. Regular audits and re-certification -- e.g., transparency -- are the keys to the long-term survivability of a repository (Online Computer Library Center, Inc. & Center for Research Libraries, 2007).

Researchers and practitioners then set about developing the criteria and checklists for audit and certification. OCLC and the Center for Research Libraries (CRL) developed the "Trustworthy Repositories Audit & Certification: Criteria and Checklist" (2007). The creators called this document "TRAC", and provided a spreadsheet for practitioners to use that covered the requirements for "organizational infrastructure", "digital object management", and "technologies, technical infrastructure, & security". The researchers with nestor (Network of Expertise in long-term STORage) also created guidelines around these three areas (Dobratz, Schoger, & Strathmann, 2006).

TRAC covered the following policy areas for audit and certification: governance & organizational viability; organizational structure & staffing; procedural accountability & policy framework; financial sustainability; contracts, licenses, & liabilities; ingest, including the creation of the archival package and acquisition of content; preservation planning; archival storage & the preservation and maintenance of AIPs; information and access management; system infrastructure; appropriate technologies; and, security (Online Computer Library Center, Inc. & Center for Research Libraries, 2007).

Ross and McHugh (2006) applied TRAC to examine mechanisms to provide audit and certification services for United Kingdom digital repositories. As part of this work, the researchers developed a toolkit, "Digital Repository Audit Method Based on Risk Assessment" (DRAMBORA). This toolkit is available online so that practitioners may "facilitate internal audit by providing repository administrators with a means to assess their capabilities, identify their weaknesses, and recognize their strengths" (Digital Curation Centre & Digital Preservation Europe, 2007). It is a self-audit that follows the workflow and criteria an external auditor would apply, so that a repository may self-assess prior to going through an external audit and certification. The toolkit provides a methodology by which a digital archivist might assess any risks to the repository she or he manages. While TRAC, DRAMBORA, and nestor are very similar, DRAMBORA provides a "documented understanding of the risks…expressed in terms of probability and impact" and provides "quantifiable insight into the severity of risks faced by repositories" along with a means to document those risks (Digital Curation Centre, 2011). In other words, TRAC is a more informal audit process that provides qualitative output, while DRAMBORA is a more detailed, formal, audit method that provides quantifiable results. The policies and risks covered by the DRAMBORA risk assessment are similar to the ones stated above for nestor and TRAC. The difference, to reiterate, is that the DRAMBORA method provides quantifiable output.

The next logical step in the development of an overall standard for the audit and certification of repositories was to merge the concepts and ideas behind TRAC, nestor, and DRAMBORA. Thus, representatives of The Digital Curation Centre (U.K.), DigitalPreservationEurope, NESTOR (Germany), and the Center for Research Libraries (North America) convened at that Chicago, IL offices of the Center for Research Libraries "to seek consensus on core criteria for digital preservation repositories, to guide further international efforts on auditing and certifying repositories" (Center for Research Libraries, 2007). Dale (2007) compared and contrasted the different methods, and created a matrix that displayed similarities and differences. Based on this matrix, and internal discussions, the attendees identified 10 core characteristics of a preservation repository:

- The repository commits to continuing maintenance of digital objects for identified community/communities.
- Demonstrates organizational fitness (including financial, staffing, and processes) to fulfill its commitment.
- Acquires and maintains requisite contractual and legal rights and fulfills responsibilities.
- Has an effective and efficient policy framework.
- Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.
- Maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.
- Creates and maintains requisite metadata about actions taken on digital objects during preservation as well as about the relevant production, access support, and usage process contexts before preservation.
- Fulfills requisite dissemination requirements.
- Has a strategic program for preservation planning and action.
- Has technical infrastructure adequate to continuing maintenance and security of its digital objects (Center for Research Libraries, 2007).

A key idea to come out of this meeting is that the preservation activities must scale to the "needs and means of the defined community or set of communities" (Center for Research Libraries, 2007). In other words, some repositories may need to implement more preservation activities, and some may need to implement less.

The Consultative Committee for Space Data Systems released the "Magenta Book" version of the "Audit and Certification of Trustworthy Digital Repositories Recommended Practice" in September 2011. This recommendation is the culmination of years of best practice work by researchers and practitioners. This best practice work began with the development of TRAC, DRAMBORA, and nestor, among other projects. The CCSDS began the process to use these methods for audit and certification to create an ISO standard, based primarily on TRAC. The precursor to this ISO standard is the

"Recommended Practice" for the "Audit and Certification of Trustworthy Digital Repositories" that is currently in release as the “Magenta Book”.

A "Recommended Practice" is not binding to any agency. The purpose of a "Recommended Practice" is to "provide general guidance about how to approach a particular problem associated with space mission support" and to provide a basis on which a community that has a stake in a digital repository may assess the trustworthiness of the repository (Consultative Committee for Space Data Systems, 2011). The CCSDS' recommendations are aimed at any and all digital repositories. Another way to think of the purpose and scope of the "Recommended Practice" is that it establishes a method for a Designated Community to determine whether or not a repository of interest is actually OAIS-compliant. The following is a summary of this Recommended Practice.

The Recommended Practice covers audit and certification criteria, including defining a "trustworthy digital repository", an evidence metric (e.g., "examples") in support of a particular requirement, and related relevant standards, best practices, and controls. The policies required to be trustworthy fall under three primary categories: "Organizational Infrastructure", "Digital Object Management", and "Infrastructure and Security Risk Management". The authors designed the document so that each of those sections follows a similar design.

First, the policy is stated. Second, the "Supporting Text" is presented; this is the "so what?" section. Third, the document provides "Examples of the Ways the Repository Can Demonstrate It Is Meeting This Requirement". Finally, the authors provide a

"Discussion" section that explains the previous three sections in order to remove any possible ambiguity.

So, for example, in section "3 Organizational Infrastructure", "3.1 Governance and Organizational Viability", section 3.1.1 states:

> **3.1.1** The repository shall have a mission statement that reflects a commitment to the preservation of, long-term retention of, management of, and access to digital information.
>
> **Supporting Text**
>
> This is necessary in order to ensure commitment to preservation, retention, management and access at the repository's highest administrative level.
>
> **Examples of Ways the Repository Can Demonstrate It Is Meeting This Requirement**
>
> Mission statement or charter of the repository or its parent organization that specifically addresses or implicitly calls for the preservation of information and/or other resources under its purview; a legal, statutory, or government regulatory mandate applicable to the repository that specifically addresses or implicitly requires the preservation, retention, management and access to information and/or other resources under its purview.
>
> **Discussion**
>
> The repository's or its parent organization's mission statement should explicitly address preservation. If preservation is not among the primary purposes of an organization that houses a digital repository then preservation may not be essential to the organization's mission. In some instances a repository pursues its preservation mission as an outgrowth of the larger goals of an organization in which it is housed, such as a university or a government agency, and its narrower mission may be formalized through policies explicitly adopted and approved by the larger organization. Government agencies and other organizations may have legal mandates that require they preserve materials, in which case these mandates can be substituted for mission statements, as they define the purpose of the organization. Mission statements should be kept up to date and continue to reflect the common goals and practices for preservation (CCSDS, 2011).

The policy areas covered by the Recommended Practice include: governance and organizational viability; organizational structure and staffing; procedural accountability and preservation policy framework; financial sustainability; contracts, licenses, and liabilities; ingest, including acquisition of content and creation of the AIP; preservation planning; AIP preservation; information management; access management;

and, risk management, including technical infrastructure and security. These areas almost exactly replicate the original audit and certification criteria for the TRAC checklist, and they are also closely replicate the criteria used in nestor and DRAMBORA.

## APPLICATIONS OF THE AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES RECOMMENDED PRACTICE AND THE OAIS REFERENCE MODEL

A complete listing of all projects, repository designs, and organizations that have applied the OAIS reference model and some version of TRAC, DRAMBORA, or nestor is beyond the scope of this literature review. Instead, this section will discuss a few example applications for applying both TRAC and the OAIS Reference Model.

When Steinhart, Dietrich, and Green (2009) applied the TRAC checklist to a "data staging repository", they made several observations and conclusions. First, the TRAC checklist is applicable "to the pilot phase of a staging repository", which is a "transitory curation environment" (Steinhart, Dietrich, & Green, 2009). This meant that TRAC had practical applications beyond digital preservation audit and certification.

For example, the TRAC checklist may be used as an evaluation tool when repository owners want to purchase new repository software. The TRAC checklist may also be used as a standard from which to create machine-actionable rules, per Smith and Moore's (2006) work on the PLEDGE project. By implementing TRAC policies at the machine-level, the amount of human effort required to enforce a policy is reduced because policy enforcement is built into the system itself (Moore & Smith, 2007).

Steinhart, Dietrich, and Green (2009) noted that there seemed to be two applications of TRAC: an audit of the system to satisfy auditors, or an audit of the system to satisfy users of the system (i.e., "the Community of Practice"). Implied in this observation is the idea that few audits seem to be conducted purely for an organization's internal erudition. Regardless of the purpose for conducting the audit, however, TRAC has provided a method for repository owners to identify gaps in an organization's workflows and policies, and provides the mechanisms (e.g., "knowledge") for those owners to fill those gaps.

Another example of the application of TRAC to a repository is the audit of the MetaArchive repository. Contractor Matt Schultz conducted an audit of the MetaArchive Cooperative and made the results public. The author reported that the MetaArchive "conforms to all 84 criteria specified by TRAC" and "has undertaken 15 reviews and/or improvements to its documentation and operations as a result of its self-assessment findings" (Educopia Institute, 2010). The organization made the actual spreadsheet available that contained the results of the audit and certification of the MetaArchive.

A quick skim of titles containing the word, "TRAC" in journals such as D-Lib Magazine, JASIST, and other related journals indicate that TRAC has been used often as an assessment tool. What is missing, however, are papers with negative assessments of TRAC, or any negative results from applying TRAC. What is also missing is a formal assessment as to whether or not a top-down approach (e.g., formal established standards) is the most feasible, or, even, the only approach. Perhaps a bottom-up approach where someone analyzes what policies people are actually implementing versus what is recommended, would be a useful approach?

Perhaps the positive reviews of the application of TRAC, of which the two above are only a small portion, indicate that as a Recommended Practice it is, indeed, comprehensive and covers all required bases. Or the positive reviews of applying TRAC may reflect researchers' and publishers' biases towards not publishing negative results, otherwise known as the "file-drawer effect" (Fanelli, 2011). The likely answer to the lack of published critical reviews of TRAC and the Recommended Practice is that not enough time has gone by to evaluate whether or not following the recommended policies will make any difference in the longevity of the repository.

As stated previously, ICPSR employees have been migrating their social science archive forward since the 1960s, with no standards such as TRAC or the OAIS to follow (Vardigan & Whiteman, 2007). Other repositories disappeared or lost information. Would having international standards in place both for repository design and audit and certification policies really have prevented that kind of loss of information? It is hard to say, as even the authors of the OAIS Reference Model state that the long-term survival of a repository depends on the will and resources of the repository owners and the community of practice (CCSDS, 2002).

Archivists and librarians at Tufts and Yale applied the OAIS Reference Model to electronic records. Specifically, they created an ingest guide to aid in moving electronic records from a "recordkeeping system to a preservation system". The practitioners designed the guide to describe the actions needed for a "trustworthy" ingest process.

The authors used both the OAIS Reference Model and the "Producer-archive Interface Methodology Abstract Standard" (Consultative Committee for Space Data Systems, 2004) as the basis for the guide. According to the archivists and librarians who worked on the project, following the guide should allow "a reasonable person to presume that a record has maintained its level of authenticity during ingest" (Fedora and the Preservation of University Records Project, 2006).

The authors divided the ingest guide into two main sections: "negotiate submission agreement" and "transfer and validation". The former section covers establishing a relationship with the collection owner, defining the project, assessing the records themselves, and finalizing the submission agreement. The latter section includes creating and transferring Submission Information Packages (SIPS), validation, transformation, metadata, formulating and assessing Archival Information Packages (AIPs), and formal accession. Each section contains an overview, an image of the flow of steps involved in that particular process, and a step-by-step written narrative for each step in the flow. The purpose of the document is not to provide "a detailed manual of procedures", but to provide "a prescriptive guide for a trustworthy ingest process" (Fedora and the Preservation of University Records Project, 2006).

A different kind of application of both TRAC and the OAIS is to build or use a "trusted digital repository" to create "persistent collections" in a "persistent archive" (Moore, 2004). Some of these solutions are based on digital library systems such as DSPACE and FEDORA; other solutions include data grids such as the Storage Resource Broker (SRB) and the integrated Rule-Oriented Data System (iRODS) (Moore, 2005; Moore, Rajasekar, & Marciano, 2007). One unique aspect of iRODS is that preservation policies outlined in TRAC may be implemented at the machine level, in the code, via the use of rules. Rajasekar, et al (2006) call this "policy virtualization".

For example, the following "human language example" regarding “Chain of Custody” from the Audit and Certification of Trustworthy Digital Repositories Recommended Practice (CCSDS, 2011):

> 5.1.2 The repository shall manage the number and location of copies of all digital objects. This is necessary in order to assert that the repository is providing an authentic copy of a particular digital object.

may be written in machine language in iRODS v.3.0 as:

```
myTestRule {
#Input parameters are:
#  Object identifier
#  Buffer for results
#Output parameter is:
#  Status
  msiSplitPath(*Path, *Coll, *File);
  msiExecStrCondQuery("SELECT DATA_ID where COLL_NAME = '*Coll' and
DATA_NAME = '*File'",*QOut);
  foreach(*QOut) {
    msiGetValByKey(*QOut,"DATA_ID",*Objid);
    msiGetAuditTrailInfoByObjectID(*Objid,*Buf,*Status);
    writeBytesBuf("stdout",*Buf);
  }
}
INPUT *Path="/tempZone/home/rods/sub1/foo1"
OUTPUT ruleExecOut
```

This type of policy virtualization is the method by which the researchers who created iRODS implemented the OAIS Reference Model recommendations within the system architecture itself (Ward, de Torcy, Chua, & Crabtree, 2009).

## OTHER TECHNICAL ISSUES

The other end of the digital curation spectrum from the OAIS Reference Model and the Audit and Certification of Trustworthy Repositories is bit level preservation. Moore (2002) wrote, "the challenge in digital archiving and preservation is not the management of the bits comprising the digital entities, but the maintenance of the infrastructure required to manipulate and display the image of reality that the digital entity represents". Lynch (2000) also writes that infrastructure is key. However, since bit level preservation is followed by preservation of the media that contains the bits and bytes, which requires preservation of the software and hardware on which the media runs, which requires networked infrastructure, bit management must be addressed.

A bit is a "binary digit". A "binary digit" is either a one or a zero in a binary system of notation (Binary digit, 2011). Chunks of bits make up a byte. Rothenberg (1999) writes that bytes may be any length, but 8 bytes provides considerably more freedom to create upper and lower case characters, punctuation, digits, control characters, and graphical elements. In very simple terms, to read a bit stream, the computer hardware must retrieve it from the media it is stored on (e.g., flash drive, CD, DVD, computer hard drive, etc.) and interpret it via software that is designed to render bits stored in that format (e.g., .pdf, .doc, .jpg., etc.).

If the bits become corrupted, then the content is unrenderable. If the media storage device deteriorates, the content is unrenderable. If the software and hardware are unavailable to read and render the file format, it is unrenderable. If the file format is unknown, then the content is unrenderable by any machine or available software. Thus, when a repository owner designs a preservation system to provide access to the content for the indefinite near-term, a decision must be made regarding migrating, refreshing, replicating, and emulating the file format, software, and hardware used to store and render the contents of a digital object.

Waters and Garrett (1996) describe migration as the transfer of data to a new operating system, programming code, or file format. The advantage of this method is it keeps the data current with technological changes. The disadvantage is that it is possible the rendering of the content may change, so that the representation is different in some way from the original (Rothenberg, 1999). In most instances, this is likely not to matter, but in some instances, it could be important. One way around this is to save the original files, migrate copies of those files to the new format/operating system/programming language, and then store the originals with the copies. The disadvantage to this, however, is that one must also save the hardware and software to read these files, which negates the advantages inherent in migration. Preservationists prefer migration to refreshing because it better retains the ability to retrieve, display and otherwise access the data (Research Libraries Group, 1996)

Archivists "refresh" data by copying data from old media onto new media, in an effort to stave off the effects media deterioration. However, this preservation method only works so long as the data and information are "encoded in a format that is independent of the particular hardware and software needed to use it and as long as there exists software to manipulate the format in current use" (Waters & Garrett, 1996). That is, the software and hardware used to read the information on the media must be backwards compatible and interoperable with different file formats, hardware, and software.

Rothenberg (1999) proposes emulation as the best solution to preservation. He defines emulation as a new system that replicates the functionality of a now-obsolete system, providing the user with the data, information, and functionality of the original system. Rothenberg writes that emulators may be built for hardware platforms, applications, and/or operating systems. However, emulation is expensive, as the cost of replicating the original system and actually being able to provide all of the functionality requires a great deal of resources, both human, financial, and time. Oltmans (2005) compared migration and emulation and concluded that emulation is more cost effective because it preserves the collection in its entirety when compared to migration. One could argue that preservationists would be better off simply maintaining the original system in the first place. However, few organizations or people have the resources to maintain that amount of hardware and software. Video game aficionados prefer to use emulators; otherwise, migration has been the method of choice for curators and preservationists of data.

Repository owners use replication as a way to back up data in multiple locations, preferable not in the same geographic or physical space. This prevents the accidental and permanent loss of data. If there is a fire, a flood, or a malicious act by some person to destroy the data, replication ensures that there are still copies of the data stored in a format such that a full restore is possible. Generally, repository managers create two replications of data. Often, this can be done in a shared format, so that one repository owner stores back up data for another organization, and vice versa. One challenge to replication is ensuring that all data stored in all locations are synced so that additions, deletions, updates, etc. are done so that the data in one location "matches" the data stored in the other two locations. The repository systems administrators much check the data on a regular basis for to ensure its continued integrity via tools such as fixity checks, access controls, and other data integrity techniques and mechanisms (Sivathanu, Wright, & Zadok, 2005). Software such as LOCKSS ("Lots of Copies Keep Stuff Safe") and data grid "middleware" such as iRODS provide repository owners with proven technology to aid in the replication of their data (Moore & Merzky, 2003; Moore, 2004; Moore, 2006). Organizations such as Data-PASS ("The Data Preservation Alliance for the Social Sciences") help their members replicate and preserve social science data by creating a common technical mechanism for data sharing/replication.

## GENERAL DIGITAL REPOSITORY MANAGEMENT

How does an archivist, librarian, or other technologist manage a preservation digital repository? The same way personnel manage a non-preservation digital repository (Lavoie & Dempsey, 2004). Material must be selected and ingested or digitized if it has not been born digital. Metadata must be created, or the quality of the metadata must be checked prior to ingest and possibly augmented if it does not meet the repository owner's quality standards (Lavoie & Gartner, 2005; Shreeves, et al, 2005; Jackson, et al, 2008; Ward, 2004). The digitization and/or ingest project must be managed, and risks to the repository must be identified and solutions created (Lawrence, et al, 2000). Intellectual property and copyright to the data must be established and enforced internally and with the Community of Interest (National Initiative for a Networked Cultural Heritage, 2002). Lee, Tibbo, & Schaefer (2007) note that the manager of the repository also must hire trained personnel with the appropriate skill sets to create, manage, preserve, and curate the repository.

## FUNDING

Last, but not least, the repository manager and the Community of Interest must ensure funding is available to maintain the repository over the indefinite long-term. And, should this funding fall short, the repository manager must ensure that there is a back up organization to take over the management of the repository, should the "owning" organization no longer exist (Waters & Garrett, 1996).

Both the members of AAMPAS (2007) and Waters & Garrett (1996) examined cost factors of preserving digital information over time. The AAMPAS members estimated the costs of digital video vs. film masters preservation, and Waters & Garrett examined digital book vs. paper book preservation and storage. Both groups reached the same conclusion: the curation and preservation of digital material is far more expensive than preserving and maintaining film or paper books over time. The AAMPAS committee determined that it would cost 1,100% more to store digital movie masters for 100 years than to store film masters for the same time period. Waters & Garrett's (1996) cost model indicated that "storage costs…are 12 times higher for a digital archives composed of texts in image form, and the access costs are 50% higher" than for the same material as books. Chapman (2003) pondered the storage affordability question and concluded that the final costs are variable. He wrote that the true costs depend on the services provided around the repository, the type and amount of content, the choice of repository software, and the type of storage chosen ("dark archive", publicly accessible, etc.).

Regardless, the final conclusion is that digital curation and preservation is not cheap. The members of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2008) noted that "there is no general agreement" as to "who is responsible…and who should pay for the access to, and preservation of, valuable present and future digital information".

## CURRENT STATUS AND FUTURE CHALLENGES/FURTHER WORK

Librarians, archivists, computer scientists and other researchers are currently immersed in figuring out the "data deluge". How big is this deluge? It is hard to estimate, but IDC estimates that the annual compound growth rate of data created and stored was almost 60% higher in 2008 than the 180 exabytes that existed in 2006 (Mearian, 2008). IDC further estimates that by 2011, "there will be 1,800 exabytes of electronic data in existence". If those numbers are correct, Mearian (2008) writes that as of 2011 the number of bits stored exceeds the number of stars in the sky.

That is a lot of data.

Digital preservationists and domain scientists are now focusing their attention on access to research data, specifically, the preservation of research data sets from which research conclusions are drawn. The Committee on Ensuring the Utility and Integrity of Research Data in the Digital Age (2009), the National Science Foundation (2005), and other individuals and groups have drawn attention to the need to steward data for use and re-use by other researchers. As one part of this, members of these two organizations have recommended creating formal standards and strategies for data stewardship.

The editors of the journal *Nature* have participated in this effort, by drawing attention to the perils and advantages of data sharing (Butler, 2007; Butler, 2007; Nelson, 2009) and data neglect (Editor, 2009). The editors of *Science* have also followed suit, and examined data sharing and data restoration (Curry, 2011; Hanson, Sugden, & Alberts, 2011) Over the course of the past year, both the National Science Foundation and the National Institutes of Health have required grant applicants to provide data management plans as part of the application process. One can only wonder at how well researchers' data management plans conform to established best practice recommendations for the preservation of data, such as the OAIS Reference Model and the Audit and Certification of Trustworthy Digital Repositories Recommended Practice.

The logic behind the interest in preserving, accessing, and sharing data sets is twofold: to ensure that science can be replicated (and the science cannot be replicated if the original data set is lost or unavailable); and to ensure that taxpayers receive the full benefits of their investment in research by allowing other researchers access to data generated with taxpayer money. If stakeholders wish to share data, then it must be stewarded when the data is gathered, on through the initial research, and includes storage of the data set(s) post-dissemination of any results. It must also be stored for the indefinite long-term, should a future researcher wish to access the data set(s).

Practitioners' initial research into this area indicates that some kind of institutional support in the form of data centers where researchers may store and share their data may be required in some instances (Beagrie, Beagrie, & Rowlands, 2009; Research Information Network, 2011). Skinner & Walters (2011) advocate a new role for librarians and archivists -- that of data curator. Their recommendation is that academic and research librarians should provide curatorial guidance with regards to digital content. They write that librarians and archivists should go to the researchers, rather than wait for the researchers to come to them for advice. Most academic and research libraries and archives do offer research data management advice, including a "data curation toolkit", to aid in interviewing the researcher about their data curation requirements (Witt, Carlson, & Brandt, 2009).

## CONCLUSION

The problem of preserving data, information, knowledge, and wisdom is not a new problem. Whether it is clay tablets, papyrus, books, data or some other format, the people who are interested in preserving the cultural, research, and other heritage of our world on earth have faced challenges of one sort or another. Some data has been preserved for centuries, and others, unnecessarily lost. War, weather, politics, fire, and other factors have destroyed valuable information objects in all centuries. The value of the data to one or more individuals is a major factor that leads to its curation and long-term survivability. The ability of the owner and users of it to fund its preservation is equally important. Librarians, archivists, and computer scientists’ establishment of standards for digital preservation and curation aid in the survivability of this data, but do not “cause” it. What has changed over time is the type of data preserved and the method for doing so. What has not changed over the millennia is that the preservation and curation of objects is not guaranteed and it is not cheap.

## REFERENCES

Ackoff, R.L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16(1), 3-9.

Beagrie, N., Beagrie, R., & Rowlands, I. (2009). Research data preservation and access: the views of researchers. *Ariadne*, 60. Retrieved August 18, 2009, from http://www.ariadne.ac.uk/issue60/beagrie-et-al/

Berners-Lee, T. (1998). *Cool URIs don't change*. W3C. Retrieved July 15, 2008, from http://www.w3.org/Provider/Style/URI.html

Binary digit. (2011). *Google.com*. Retrieved December 13, 2011, from http://www.google.com/search?client=safari&rls=en&q=define:+binary+digit&ie=UTF-8&oe=UTF-8

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2008, December). *Sustaining the digital investment: issues and challenges of economically sustainable digital preservation*. San Diego, CA: San Diego Supercomputer Center. Retrieved January 24, 2009, from http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

Brody, T. (2000). *Mining the social life of an ePrint archive*. Retrieved September 16, 2001, from the University of Southampton, OpCit Project Web site: http://opcit.eprints.org/tdb198/opcit/q2/

Buckland, M.K. (1991). Information as thing. *Journal of the American Society for Information Science*, 42(5), 351-360.

Butler, D. (2007). Agencies join forces to share data. *Nature*, 446, 354.

Butler, D. (2007). Data sharing: the next generation. *Nature*, 446, 10-11.

Carr, L. (1999). *Metadata changes to XXX papers in a three month period*. Retrieved October 13, 2001, from the University of Southampton, Electronics and Computer Science Web site: http://users.ecs.soton.ac.uk/lac/XXXmetadatadeltas.html

Center for Research Libraries (2007). *Ten principles*. Retrieved December 8, 2011, from http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re

Committee on Ensuring the Utility and Integrity of Research Data in the Digital Age; National Academy of Sciences. (2009). *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Executive Summary. Washington, DC:

the National Academies Press. Retrieved January 7, 2009, from http://www.nap.edu/catalog.php?record_id=12615

CCSDS. (2002). *Reference model for an Open Archival Information System (OAIS)* (CCSDS 650.0-B-1). Washington, DC: National Aeronautics and Space Administration (NASA). Retrieved April 3, 2007, from http://nost.gsfc.nasa.gov/isoas/

CCSDS. (2004). *Producer-archive interface methodology abstract standard* (CCSDS 651.0-B-1). Washington, DC: National Aeronautics and Space Administration (NASA). Retrieved August 18, 2007, from http://public.ccsds.org/publications/archive/651x0b1.pdf

CCSDS. (2011). *Audit and Certification of Trustworthy Digital Repositories* (CCSDS 652.0-M-1). Magenta Book, September 2011. Washington, DC: National Aeronautics and Space Administration (NASA).

Curry, A. (2011). Rescue of Old Data Offers Lesson for Particle Physicists. *Science*, 331, 694-695.

Dale, R. (2007). *Mapping of audit & certification criteria for CRL meeting (15-16 January 2007)*. Retrieved September 11, 2007, from http://wiki.digitalrepositoryauditandcertification.org/pub/Main/ReferenceInputDocuments/TRAC-Nestor-DCC-criteria_mapping.doc

Davis, J. R. & Lagoze, C. (2000). NCSTRL: design and deployment of a globally distributed digital library. *Journal of the American Society for Information Science*, 51(3), 273-280.

Digital Curation Centre. (2011). *DRAMBORA*. Retrieved December 9, 2011, from http://www.dcc.ac.uk/resources/tools-and-applications/drambora

Digital Curation Centre. (2010). *What is digital curation*? Retrieved November 6, 2011, from http://www.dcc.ac.uk/digital-curation/what-digital-curation

Digital Curation Centre & Digital Preservation Europe. (2007). *DCC and DPE digital repository audit method based on risk assessment* (DRAMBORA). Retrieved August 1, 2007, from http://www.repositoryaudit.eu/download

Digital Preservation. (2009). *Introduction - definitions and concepts*. Digital Preservation Coalition. Retrieved November 6, 2011, from http://dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

Dobratz, S., Schoger, A., & Strathmann, S. (2006). *The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification*. Paper presented at the workshop on “digital curation & trusted repositories: seeking success”, held in conjunction with the ACM/IEEE Joint Conference on Digital Libraries, June 11-15, 2006, Chapel Hill, NC, USA. Retrieved December 1, 2011, from http://www.ils.unc.edu/tibbo/JCDL2006/Dobratz-JCDLWorkshop2006.pdf

Duranti, L. (1995). Reliability and authenticity: the concepts and their implications. *Archivaria*, 39 (Spring), 5-10.

Editor. (2009). Data's shameful neglect. *Nature*, 461, 145.

Educopia Institute. (2010, April). *Metaarchive cooperative TRAC audit checklist*. Prepared by M. Schultz. Atlanta, CA: Educopia Institute. Retrieved December 10, 2010 from http://www.metaarchive.org/sites/default/files/MetaArchive_TRAC_Checklist.pdf

Egger, A. (2006). Shortcomings of the Reference Model for an Open Archival Information System (OAIS). *IEEE TCDL Bulletin*, 2(2). Retrieved October 23, 2009, from http://www.ieee-tcdl.org/Bulletin/v2n2/egger/egger.html

Fedora and the Preservation of University Records Project. (2006). *2.1 Ingest Guide, Version 1.0* (tufts:central:dca:UA069:UA069.004.001.00006). Retrieved April 16, 2009, from the Tufts University, Digital Collections and Archives, Tufts Digital Library Web site: http://repository01.lib.tufts.edu:8080/fedora/get/tufts:UA069.004.001.00006/bdef:TuftsPDF/getPDF

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, September 2011, 1-14.

Galloway, P. (2004). Preservation of digital objects. In B. Cronin (Ed.), *Annual Review of Information Science and Technology*, 38(1), (pp. 549-590).

Ginsparg, P. (2011). ArXiv at 20. *Nature*, 476, 145-147.

Hanson, B., Sugden, A., & Alberts, B. (2011). Making Data Maximally Available. *Science*, 331, 649.

Hedstrom, M. (1995). Electronic archives: integrity and access in the network environment. *American Archivist*, 58(3), 312-324.

Higgens, S. (2007). Draft DCC curation lifecycle model. *International Journal of Digital Curation*, 2(2). Retrieved March 22, 2008, from http://www.ijdc.net/index.php/ijdc/article/view/46

InterPARES. (2001). *The long-term preservation of authentic electronic records: findings of the InterPARES project*. Retrieved October 5, 2007, from http://www.interpares.org/ip1/ip1_index.cfm

Jackson, A. S., Han, M., Groetsch, K., Mustafoff, M., & Cole, T. W. (2008). Dublin Core metadata harvested through the OAI-PMH (pre-print). *Journal of Library Metadata*, 8(1).

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2), 162-180.

Koehler, W. (2004). A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2).

Krasner-Khait, B. (2001). Survivor: the history of the library. *History Magazine*, October/November 2011. Retrieved August 30, 2011, from http://www.history-magazine.com/libraries.html

Kunze, J. (2003). *Towards electronic persistence using ARK identifiers*. Retrieved July 10, 2008, from the University of California, California Digital Library, Inside CDL Web site: http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf

Lagoze, C. and Van de Sompel, H. (2001). The Open Archives Initiative: building a low-barrier interoperability framework. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, June 24-28, 2001, Roanoke, VA. pp. 54-62.

Lavoie, B. (2004). *The open archival information system reference model: introductory guide*. Technology Watch Report. Dublin, OH: Digital Preservation Coalition. Retrieved March 6, 2007, http://www.dpconline.org/docs/lavoie_OAIS.pdf

Lavoie, B. & Dempsey, L. (2004). Thirteen ways of looking at...digital preservation. *D-Lib Magazine*, 10(7/8). Retrieved May 7, 2007, from http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

Lavoie, B. & Gartner, R. (2005). *Preservation metadata*. Technology Watch Report. Dublin, OH: Digital Preservation Coalition. Retrieved June 20, 2009, http://www.dpconline.org/docs/reports/dpctw05-01.pdf

Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. (2000). *Risk management of digital information: a file format investigation*. Washington, DC: Council on Library and Information Resources. Retrieved October 22, 2007, from http://www.clir.org/pubs/reports/pub93/contents.html

Lee, C. (2010). Open archival information system (OAIS) reference model. In *Encyclopedia of Library and Information Sciences*, Third Edition. London: Taylor & Francis.

Lee, C., Tibbo, H.R., & Schaefer, J.C. (2007). Defining what digital curators do and what they need to know: The DigCCurr Project. In *Proceedings of the 2007 ACM/IEEE Joint Conference on Digital Libraries*, 49-50.

Lynch, C. A. (1994). The integrity of digital information: mechanics and definitional issues. *Journal of the American Society for Information Science*, 45(10), 737-744.

Lynch, C. (2000). Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust. *Authenticity in a digital environment*. Washington, DC: Council in Library and Information Resources. Retrieved April 14, 2009, from http://www.clir.org/pubs/reports/pub92/pub92.pdf

McCown, F., Chan, S., Nelson, M.L., & Bollen, J. (2005). *The availability and persistence of Web references in D-Lib Magazine*. Paper presented at the 5th International Web Archiving Workshop (IWAW05), Vienna, Austria. Retrieved July 14, 2008, from http://arxiv.org/abs/cs.OH/0511077

Mearian, L. (2008). Study: digital universe and its impact bigger than we thought. *Computerworld*, March 11, 2008. Retrieved March 14, 2008, from http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9067639

Moore, R. (2002). The preservation of data, information, and knowledge. In *Proceedings of the World Library Summit*, April 24-26, 2002, Singapore. Retrieved April 1, 2009, from http://www.sdsc.edu/NARA/Publications/Web/moore-rw.doc

Moore, R. (2004). *Evolution of data grid concepts*. Paper presented at the workshop on “data” at the 10th Global Grid Forum, Berlin, Germany, March 9-13, 2004. Retrieved March 23, 2009, from http://www.npaci.edu/DICE/Pubs/Grid-evolution.doc

Moore, R.W. (2004). Preservation Environments. In *Proceedings of the NASA/IEEE MSST 2004 Twelfth NASA Goddard Conference on Mass Storage Systems and Technologies in cooperation with the Twenty-First IEEE Conference on Mass Storage Systems and Technologies (MSST 2004)*, April 13-16, 2004, Adelphi, Maryland, USA. Retrieved September 26, 2010, from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20040121020_2004117345.pdf

Moore, R. (2005). Persistent collections. In S.H. Kostow & S. Subramaniam (Eds.), *Databasing the brain: from data to knowledge (neuroinformatics)* (pp. 69-82). Hoboken, NJ: John Wiley and Sons.

Moore, R. (2006). Building preservation environments with data grid technology. *American Archivist*, 69(1), 139-158.

Moore, R. & Merzky, A. (2003). *Persistent archive concepts*. Paper presented at the 7th Global Grid Forum, Tokyo, Japan, March 4-7, 2003. Retrieved March 4, 2009, from http://www.npaci.edu/DICE/Pubs/Data-PAWG-PA.doc

Moore, R., Rajasekar, A., & Marciano, R. (2007). Implementing Trusted Digital Repositories. In *Proceedings of the DigCCurr2007 International Symposium in Digital Curation*, University of North Carolina - Chapel Hill, Chapel Hill, NC USA, 2007. Retrieved September 24, 2010, from http://www.ils.unc.edu/digccurr2007/papers/moore_paper_6-4.pdf

Moore, R. & Smith, M. (2007). Automated Validation of Trusted Digital Repository Assessment Criteria. *Journal of Digital Information*, 8(2). Retrieved March 2, 2010, from http://journals.tdl.org/jodi/article/view/198/181

National Initiative for a Networked Cultural Heritage. (2002). Rights management. In *the NINCH guide to good practice in the digital representation and management of cultural heritage materials*, v.1.0. Glasgow: University of Glasgow (HATII) & NINCH. Retrieved April 17, 2009, from http://www.nyu.edu/its/humanities/ninchguide/IV/

National Science Foundation. (2005). *Long-lived digital data collections enabling research and education in the 21st century (NSB-05-40).* Arlington, VA: National Science Foundation. Retrieved May 5, 2008, from http://www.nsf.gov/pubs/2005/nsb0540/

Nelson, B. (2009). Data sharing: empty archives. *Nature*, 461, 160-163.

Nelson, M.L. (2000). *Buckets: Smart Objects for Digital Libraries* (Doctoral Dissertation). Retrieved December 14, 2011, from http://www.cs.odu.edu/~mln/phd/

Nelson, M.L., & Allen, B.D. (2002). Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1). Retrieved July 18, 2007, from http://www.dlib.org/dlib/january02/nelson/01nelson.html

NESTOR Working Group on Trusted Repository — Certification. (2006). *Catalog of criteria for trusted digital repositories version 1 draft for public comment* (urn:nbn:de:0008-2006060703). Berlin: nestor Working Group — Certification. Retrieved April 14, 2009, http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf

Oltmans, E. & Kol, N. (2005). A comparison between migration and emulation in terms of costs. *RLG DigiNews* 9(2). Retrieved September 10, 2007, from http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file959.html

Online Computer Library Center, Inc. & Center for Research Libraries. (2007). *Trustworthy repositories audit & certification: criteria and checklist version 1.0*. Dublin, OH & Chicago, IL: OCLC & CRL. Retrieved September 11, 2007, from http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

O'Toole, J.M. (1989). On the idea of permanence. *American Archivist*, 52, 10-25.

Paskin, N. (2003). DOI: A 2003 progress report. *D-Lib Magazine*, 9(6). Retrieved July 9, 2008, from http://www.dlib.org/dlib/june03/paskin/06paskin.html

Rajasekar, A., Wan, M., Moore, R., & Schroeder, W. (2006). *A prototype rule-based distributed data management system*. Paper presented at a workshop on "next generation distributed data management" at the High Performance Distributed Computing Conference, June 19-23, 2006, Paris, France.

Research Information Network. (2011). *Data centres: their use, value, and impact*. A Research Information Network report. London, UK: JISC, September 2011.

Research Libraries Group. (1996)*. Preserving digital information report of the task force on archiving of digital information*. Final report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group. Retrieved September 24, 2007, from http://www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED395602&ERICExtSearch_SearchType_0=eric_accno&accno=ED395602

Research Libraries Group. (2002). *Trusted digital repositories: attributes and responsibilities an RLG-OCLC report*. Mountain View, CA: Research Libraries Group. Retrieved September 11, 2007, from http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

Rosenthal, D.S.H., Robertson, T., Lipkis, T., Reich, V., Morabito, S. (2005). Requirements for digital preservation systems a bottom-up approach. *D-Lib Magazine*, 11(11). Retrieved August 11, 2007, from http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

Ross, S. & McHugh, A. (2006). The role of evidence in establishing trust in repositories. *D-Lib Magazine* 12(7/8). Retrieved May 6, 2007, from http://www.dlib.org/dlib/july06/ross/07ross.html

Rothenberg, J. (1999). *Avoiding technological quicksand: finding a viable technical foundation for digital preservation* (pub 77). A report to the Council on Library and Information Resources. Washington, DC: Council on Library and Information Resources. Retrieved April 16, 2009, from http://www.clir.org/pubs/reports/rothenberg/pub77.pdf

Rothenberg, J. (1999). *Ensuring the longevity of digital information*. Washington, DC: Council on Library and Information Resources. Retrieved April 16, 2009, from http://www.clir.org/pubs/archives/ensuring.pdf

Society of American Archivists. (1999). Core Archival Functions. *Guidelines for College and University Archives*. Prepared by the College and University Archives Section of the Society of American Archivists (SAA). Retrieved May 26, 2010, from http://www.archivists.org/governance/guidelines/cu_guidelines4.asp

Science and Technology Council. (2007). *The digital dilemma strategic issues in archiving and accessing digital motion picture materials*. The Science and Technology Council of the Academy of Motion Picture Arts and Sciences. Hollywood, CA: Academy of Motion Picture Arts and Sciences.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is 'quality' metadata 'shareable' metadata? The implications of local metadata practices for federated collections. In *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries*, April 7-10 2005, Minneapolis, MN, 223-237.

Sivathanu, G., Wright, C.P., & Zadok, E. (2005). Ensuring data integrity in storage: techniques and applications. In *Proceedings of the first ACM International Workshop on Storage Security and Survivability (StorageSS 05),* held in conjunction with the 12th ACM Conference on Computer and Communications Security (CCS 2005), November 7-11, 2005, Alexandria, VA. Retrieved October 4, 2007, from http://www.fsl.cs.sunysb.edu/docs/integrity-storagess05/integrity.html

Smith, M. & Moore, R. (2006). *Digital Archive Policies and Trusted Digital Repositories*. Paper presented at the 2nd International Digital Curation Conference, November 21 - 22, 2006, Glasgow, Scotland. Retrieved November 2, 2009, from http://pledge.mit.edu/images/6/6f/Smith-Moore-DCC-Nov-2006.pdf

Steinhart, G., Dietrich, D., & Green, A. (2009). Establishing trust in a chain of preservation the TRAC checklist applied to a data staging repository (DataStaR). *D-Lib*

*Magazine* 15(9/10). Retrieved October 13, 2009 from http://www.dlib.org/dlib/september09/steinhart/09steinhart.html

Thibodeau, Kenneth. (2002). Overview of technological approaches to digital preservation and challenges in coming years. In *Proceedings of the State of Digital Preservation: An International Perspective*, at the Institutes for Information Science, April 24-25, 2002, Washington, DC. Retrieved September 26, 2007 from http://www.clir.org/pubs/reports/pub107/thibodeau.html

Thibodeau, K. (2007). The Electronic Records Archives Program at the National Archives and Records Administration. *First Monday*, 12(7). Retrieved January 15, 2009 from http://firstmonday.org/issues/issue12_7/thibodeau/index.html

Tibbo, H.R. (2003). On the nature and importance of archiving in the digital age. *Advances in Computers*, 57, 1-67.

URI Planning Interest Group. (2001). *URIs, URLs, and URNs: Clarifications and Recommendations 1.0*. Report from the joint W3C/IETF URI Planning Interest Group, W3C Note, 21 September 2001. Retrieved November, 8, 2011, from http://www.w3.org/TR/uri-clarification/

Vardigan, M. & Whiteman, C. (2007). *ICPSR meets OAIS: applying the OAIS reference model to the social science archive context*. Archival Science, 7(1). Netherlands: Springer. Retrieved February 20, 2008, from http://www.springerlink.com/content/50746212r6g21326/

Walters, T. & Skinner, K. (2011). *New roles for new times: digital curation for preservation*. Report prepared for the Association of Research Libraries. Washington, D.C.: Association of Research Libraries. Retrieved April 2, 2011, from http://www.arl.org/bm~doc/nrnt_digital_curation17mar11.pdf.

Ward, J. (2004). Unqualified Dublin Core usage in OAI-PMH Data Providers. *OCLC Systems and Services*, 20(1), 40-47.

Ward, J.H., de Torcy, A., Chua, M., and Crabtree, J. (2009). Extracting and Ingesting DDI Metadata and Digital Objects from a Data Archive into the iRODS extension of the NARA TPAP using the OAI-PMH. In *Proceedings of the 5th IEEE International Conference on e-Science*, Oxford, UK, December 9-11, 2009.

Waters, D. and Garrett, J. (1996). *Preserving Digital Information*. Report of the Task Force on Archiving of Digital Information. Washington, DC: CLIR, May 1996.

Wells, H.G. (1938). *World brain*. Garden City, NY: Doubleday, Doran and Co.

Witt, M., Carlson, J., & Brandt, D.S. (2009). Constructing data curation profiles. *International Journal of Digital Libraries*, 3(4), 93-103.

Zen College Life. (2011). *The history of libraries through the ages*. Retrieved August 30, 2011, from http://www.zencollegelife.com/the-history-of-libraries-through-the-ages/