View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Oregon Scholars' Bank

Presented to the Interdisciplinary Studies Program:

UNIVERSITY OF OREGON
APPLIED INFORMATION MANAGEMENT

Applied Information Management
and the Graduate School of the
University of Oregon
in partial fulfillment of the
requirement for the degree of
Master of Science

Digital Data Preservation Practices to Ensure Long-Term Information Accessibility

CAPSTONE REPORT

Amir Rizk
Network Analyst
Inland Northwest Health Services

University of Oregon
Applied Information
Management
Program

July 2013

Continuing Education
1277 University of Oregon
Eugene, OR 97403-1277
(800) 824-2714

Approved by

_____

Dr. Linda F. Ettinger
Senior Academic Director, AIM Program

Digital Data Preservation Practices to Ensure Long-Term Information Accessibility

Amir Rizk

Inland Northwest Health Services

Abstract

This annotated bibliography describes data preservation practices used by organizations to ensure long-term digital content accessibility. It is intended for information professionals working in data management roles. Selected literature is published between 2001 and 2013. Practices designed to ensure data accessibility for as long as needed include: (a) an interconnected network of trustworthy repositories, (b) guidelines for data stewardship, (c) a Trustworthy Digital Object (TDO) design (encapsulation and encoding), and (d) preservation self-assessment tools for libraries.

Keywords: *data archiving, data preservation, digital archiving, information preservation, persistent archives, future accessibility, historical accessibility*.

Table of Contents

Introduction

Problem Area

Information preservation, defined by Chen (2001) as "keeping [data] unchanged" (p. 3), is a key challenge that faces information managers in all organizations and businesses. Information preservation in the analog world used to entail dealing with "static objects, grabbing all of their contents and storing them in some form" (Chen, 2001, p. 3). In the digital world, Chen (2001) notes the need to ensure data remains not only accessible, but also readable, including the ability to access the information using the available technology and tools (p. 2).

According to Lorie (2001), inquiry into information preservation in the digital world, (including digital archiving) must address the need to guarantee the information's long-term survival despite changes in storage media, devices, and data formats (p. 346). Chen (2001) notes that same year that "high end computing generates scientific and nationally critical data, which must be preserved over the long term" (p.4). Ignoring the problem and not taking steps towards ensuring preservation of digital content is "analogous to fostering cultural and intellectual poverty and squandering potential long-term gains" to the detriment of future generations (Chen, 2001, p. 2).

Unfortunately, ensuring data preservation is not always guaranteed with the rapid rate of data generation due to the use of modern digital authoring tools (Berman, 2008, p. 51). Short media life, obsolete hardware and software, and websites that no longer exist, make accessing older data hard; particularly with archiving standards easily ignored (Donaldson & Yakel, 2012, p. 56). According to Chen's (2001) observation over a decade ago, "the majority of products and services on the market today did not exist five years ago" (p. 2). More recently, Berman (2008)

concludes that the way digital data is stored, accessed, managed, and preserved are challenges that need attention (p. 51).

Purpose

The purpose of this scholarly annotated bibliography is to identify and present literature that describes digital data preservation practices (including digital archiving) to ensure long-term information accessibility (Goth, 2012, p. 13). The primary goal of this paper is to (a) highlight how selected organizations apply specific practices to ensure data is usable for as long as it needs to be kept and accessed within the organization, (b) identify the relevant best practices used to ensure data preservation, and (c) showcase the work of libraries in archiving content (Smith & Moore, 2007, p. 92). In this context, *organizations* are defined by Berman (2008) as publicly owned companies; scientific and engineering communities; and research and educational communities (p.52). Additionally, preservation practices in this study are limited to those that reflect the most current knowledge about professional practices to increase long-term accessibility, including (a) interoperability, (b) consistency, and (c) the safety and security of collections, as defined by Donaldson and Yakel (2012, p. 55).

Lorie (2001) and Careless (2013) each propose different means to maintain accessibility of data over time. For example, Lorie (2001) suggests "rejuvenating the information periodically by copying it from the old medium onto a newer one" (p. 346); however he does not think that solving the archiving problem is simply a technical challenge (p. 352). Careless (2013) advocates (a) storing the data in multiple geographically separate locations and on different media types to minimize risk and (b) refreshing the media as storage is updated to keep data accessible (p. 45). Careless (2013) also indicates that the problem remains to this day, and states that it requires a data migration plan to be incorporated into the planning of any new storage solution to address

hardware technology, file system, and file formats and their major changes along with reader and access software (p. 46) to ensure accessibility.

Audience

This annotated bibliography is designed for professionals who are responsible for managing information preservation, specifically systems administrators who are responsible for managing digital data for their organizations so that it can be successfully retrieved in the future (Moore, 2008, p. 64). These professionals administer backup systems and storage infrastructure, and decide on standards to be incorporated and set for current and future use within the organization (Careless, 2013, p. 46). As part of their work, they need to be aware of historical use of data in the organization, and accordingly have the knowledge to predict future trends (Breeding, 2013, p. 26). The assumption underlying this bibliography is that by providing a view of relevant practices designed to ensure data is usable for as long as it needs to be kept within the organization, information preservation managers will be better able to make decisions and resolve problems associated with preservation (Careless, 2013, p. 46).

Research Questions

This annotated bibliography examines what selected organizations are doing to address information preservation by identifying literature that addresses data preservation practices and digital archiving methods. The main question of the study is what are the relevant preservation practices designed to ensure data is usable for as long as it needs to be kept within the organization. The references presented in the Annotated Bibliography section of the document are organized in three sections defined by the following research sub-questions: (a) what are the relevant practices adopted by private and public organizations involved in commerce, science, and engineering for disaster recovery and long-term accessibility; (b) what best practices exist to

manage data preservation; and (c) what efforts have been adopted by libraries for the purpose of archiving content to ensure (a) interoperability, (b) consistency, and the (c) safety and security of collections (Donaldson & Yakel, 2012, p. 55).

Delimitations

Time frame. When collecting materials for this research, recent publications receive more attention than older ones to maintain currency. Accordingly, most selected references fall within the last six years (2006 – 2013). However, in some cases, older material such as Chen's, dating back to a slightly over a decade ago is used to address the evolution of the research topic (information preservation), and demonstrate the history of research in this field.

Selection criteria. Selected literature is retrieved from University of Oregon's (UO) library databases, with specific focus on the computer science databases including IEEE and ACM, the library sciences databases, EBSCO Host, and Google Scholar, as well as references from papers retrieved through the above databases. Extra attention is paid to peer-reviewed articles retrieved from academic databases.

Audience. This annotated bibliography is targeted at systems administrators who work with data, backup and archiving, and storage technologies. It may not be as relevant to upper management who are interested in end results as opposed to system design.

Topic definition. While information preservation can be defined in a number of different ways, this research focuses on the definition of preservation practices presented by Donaldson and Yakel (2013), limited to those that reflect the most current knowledge about professional practices to increase long-term accessibility including (a) interoperability, (b) consistency, and the (c) safety and security of collections.

Best practices. This annotated bibliography focuses on best practices recommended to ensure information preservation and accessibility. While there are various methods and systems used by many, this report highlights a few select practices deemed the most effective as recommended by professionals to preserve and provide access to data (Meyer, 2009, p. 24).

Reading and Organization Plan Preview

Based on the critical reading approach presented by Busch et al. (2013), the reading and analysis of the references follow these steps:

1. Determining the level of analysis.
2. Determining how many concepts to use.
3. Determining whether to focus on existence or frequency of a concept.
4. Determining how to distinguish among concepts.
5. Developing coding rules.
6. Determining what information is irrelevant.
7. Coding the text.
8. Analysis of results.

The organization plan for this annotated bibliography is based on the University of North Carolina's (2012) thematic plan. The plan is organized around three themes corresponding to the sub-questions presented in the research questions section. Each reference is previewed and categorized in the Annotated Bibliography section of this document under one of the following headers: (a) what are the relevant practices adopted by private and public organizations involved in commerce, science, and engineering for disaster recovery and long-term accessibility; (b) what best practices exist to manage data preservation; and (c) what efforts have been adopted by libraries for the purpose of archiving content to ensure (a) interoperability, (b) consistency, and

the (c) safety and security of collections (Donaldson & Yakel, 2012, p. 55). The same organizational plan is utilized to present results of the coding process in the Conclusion.

Definitions

To mitigate the risk of confusion in using terms that may have more than one meaning, the following list of definitions is provided to give the reader an insight into the scope of the paper. Readers from backgrounds outside the topic also find this useful as it explains terms whose use goes beyond common language (Creswell, 2009, p. 39). Definitions are derived from references presented in the Annotated Bibliography.

Archiving Standards – The most current knowledge about professional practices to increase interoperability, consistency, and the safety and security of collections (Donaldson & Yakel, 2012, p. 55).

Best Practices – Efforts and recommendations by professionals to determine the best methods of preserving and providing access to data (Meyer, 2009, p. 24).

Curation – Maintaining and adding value to a trusted body of digital information for current and future use (Berman, 2008, p.55).

Data Archiving / Digital Archiving – Ensuring that a copy of a preserved document survives as long as needed and authorized consumers can find and use any preserved document while ensuring trustworthiness (Gladney, 2006, p. 2).

Data Preservation – Ensuring the long term viability and availability of digital material (Berman, 2008, p. 55).

Digital Data – A machine readable representation of data in a recognizable format (Ludascher, Marciano, & Moore, 2001, p. 55).

Information Accessibility – Providing greater access to digital material and making it widely accessible (Balas, 2007, p. 32).

Organizations – Publicly owned companies, scientific and engineering communities, and research and educational communities (Berman, 2008, p. 52).

Preservation Practice – The initial decision to adopt preservation and the actual subsequent steps taken in the process where implementation actually occurs (Donaldson & Yakel, 2012, p. 57).

Storage Medium - Local server storage, portable hard drives, CDs, file servers, and storage arrays used to store data (Careless, 2013, p. 45).

Systems Administrator – A data processing person who manages programs and documents, along with their conversion and storage (Rothenberg, 1999, p. 14).

Research Parameters

This section defines the research parameters used in the design of this study. It lists information pertaining to the search strategy, key terms used in the research, the reference collection strategy, evaluation criteria of selected texts, the documentation approach of texts, and the reading and organization plans used for this annotated bibliography.

Search Strategy

The search strategy for this annotated bibliography is focused on querying and researching the University of Oregon's (UO) online databases. The UO Library Sciences databases provide a good resource for finding papers such as Katre, Donaldson, Gaur, Ray, Harvey, and others. Some of the Library Sciences databases overlap with those of Computer Science, which displays some papers already reviewed.

Most of the references gathered deal with archiving data in libraries and large organizations because these entities have put the most effort into preservation, but as noted by Berman (2008), other groups are starting to grapple with the responsibility of having to create plans for stewardship of digital data (p. 55). The digital preservation division of the Library of Congress is an important source to investigate as a resource for standards and leading initiatives.

Another source of references is bibliographies and reference lists of already found references, which provide a foundation for further research, including historical aspects that help to more thoroughly understand the evolution of the subject and relevant definitions.

Key Terms

The main search term is *data archiving*; with additional search terms including, *data preservation, digital archiving, information preservation, persistent archives, future accessibility,* and *historical accessibility*. The additional search phrases and key words

developed iteratively, as a result of tags and thesauruses associated with articles previously found.

Reference Collection

To ensure relevancy of collected references, each resource is reviewed in a similar manner to confirm equal treatment of all references and enforce the relationship to the research question (Hewitt, 2002, p. 22). The following steps are taken to analyze the content:

1. The abstract of each resource is read to determine if it is a relevant resource of value to the purpose of the research.
2. If the paper is unrelated or focuses on an aspect not relevant to the goals of this study, it is discarded.
3. If the abstract proves useful, the citation is saved, along with the source where it was found, and the abstract itself.
4. The list of collected sources is scanned again, and the general or older resources are kept for the reference list, while more current and relevant sources are filed under the appropriate sub-question for use in the annotated bibliography.
5. The relevant papers are downloaded, saved, and scanned for useful quotations.
6. Useful quotations are saved in a separate document along with the citation.

Once a source is determined to be of value to the annotated bibliography, it is downloaded and saved locally using the author's name, the publication date, and the title of the paper. The citation is then saved to the list of discovered resources and filed under the appropriate heading corresponding to one of three research sub-questions: (a) what are the relevant practices adopted by private and public organizations involved in commerce, science, and engineering for disaster recovery and long-term accessibility; (b) what best practices exist to

manage data preservation; and (c) what efforts have been adopted by libraries for the purpose of archiving content to ensure (a) interoperability, (b) consistency, and the (c) safety and security of collections (Donaldson & Yakel, 2012, p. 55). Any resources that are relevant, but do not specifically tie into any of the listed sub-questions are filed under the general header, and listed in the references section.

Evaluation Criteria

The main source of literature for this review is articles published in academic and professional journals. However, other sources are utilized to cover more content, including magazine articles and conference reports. Attention is given to the relevancy and quality of each resource to ensure selection of literature is consistent with the purpose and goals presented in this review (Bell & Frantz, 2012).

Following the guidelines provided by Bell and Frantz (2012), relevancy of sources is determined by a combination of date published and the type of publication. Published abstracts provide a quick look into the contents of each source and provide a preliminary assessment of the relevance of that source to the literature review and the topic on hand. Articles published recently in library journals addressing archiving and data accessibility are good sources to include, although some are too specific and focus on an angle not covered in this review. Such sources are not aligned with the focus of this paper and are not included. Further investigation is often required to assess the relevancy of a source, and is achieved by reading the whole paper or article to determine significance to the topic. The age of the sources is a relevant benchmark to consider, however, age alone is not the deciding factor. Some of the older resources are still relevant due to the groundwork they lay for more recent papers. Some older literature is chosen to define context, show historical use, and provide definitions of terms and concepts used.

Quality of resources is also evaluated to ensure credibility of sources. Credible sources include peer-reviewed articles published in academic and scientific journals, being cited in other credible papers, being written by an expert in the field, or being published in a relevant trade or business periodical. Objectivity in tone is also an indication of the quality of research and writing provided in a source. The format that papers are presented in also attests to the quality and credibility of a source (Bell & Frantz, 2012).

Documentation Approach

Once a reference is determined to be of value to the annotated bibliography, it is downloaded and saved locally using the author's name, the publication date, and the title of the paper. The citation is then saved to the list of key references, and filed under the appropriate heading, corresponding to one of three research sub-questions: (a) what are the relevant practices adopted by private and public organizations involved in commerce, science, and engineering for disaster recovery and long-term accessibility; (b) what best practices exist to manage data preservation; and (c) what efforts have been adopted by libraries for the purpose of archiving content to ensure (a) interoperability, (b) consistency, and the (c) safety and security of collections (Donaldson & Yakel, 2012, p. 55). References that are relevant, but do not specifically tie into any of the listed sub-questions, are listed under the general References section.

Reading and Organization Plan

Reading plan. Busch at al. (2013) present an approach to critical reading and conceptual analysis of references to examine words or phrases within the text in order to determine its usability in the research. A reading plan based on their approach is used in this annotated bibliography and listed here.

Reading a text and analyzing it starts with deciding upon the level of analysis, choosing the phrases to look for in the text, and pre-defining the selection of words or phrases. In this annotated bibliography, the coding phrases include *data archiving, data preservation, digital archiving, information preservation, persistent archives, future accessibility,* and *historical accessibility* as defined in the key terms section. Some texts present additional relevant words or phrases, and that is assessed as it occurs differently for each text. Coding then proceeds based on existence of phrases in the text to assess importance to each specific author and his or her focus within the text.

Distinguishing among the terms and concepts is also determined based on the author's use of synonyms, or if the terms are used to mean different things. At this point, a translation rule is created to streamline and organize the coding process. This guarantees consistency of coding across the different texts. These rules align with the Definitions section presented in this paper. Irrelevant information and phrases are discarded for lack of applicability.

The actual coding process proceeds by recording and compiling a list of the words and phrases from each text, including those determined to be synonyms. Once completed, the analysis ensues to draw relevant conclusions and prioritize and categorize the recorded concepts as noted in the Organization Plan below.

Organization plan. Upon finishing the reading, the references are organized and presented in the Annotated Bibliography section of this document based on themes under the appropriate headers corresponding to the research sub-questions (University of North Carolina, 2012). This approach aims to answer each of the research sub-questions by presenting literature that is relevant to the question, and can directly address its inquiry. The following themes are used to organize the references: (a) the relevant practices adopted by private and public

organizations involved in commerce, science, and engineering for disaster recovery and long-term accessibility. This theme covers existing practices used by organizations to archive data, ensuring that the data is accessible in the future for more than a decade (Hart, 2003, p. 94); (b) the best practices to manage data preservation. This theme attempts to identify specific best practices and recommendations (Meyer, 2009, p. 5) to be used to ensure preservation of data; (c) the efforts adopted by libraries for the purpose of archiving content to ensure (a) interoperability, (b) consistency, and the (c) safety and security of collections (Donaldson & Yakel, 2012, p. 55). This final theme introduces particular examples from libraries that have addressed information preservation specifically in their operations and practices.

Annotated Bibliography

The purpose of this annotated bibliography is to identify and present literature that describes digital data preservation practices (including digital archiving) to ensure long-term information accessibility. Annotations consist of four elements: (a) the bibliographic citation in APA format; (b) the published abstract; (c) a description of the credibility of the reference; and (d) a summary of the relevant content related to the research questions addressed in this study.

The main question of the study is: what are the relevant preservation practices designed to ensure data is usable for as long as it needs to be kept within the organization. The references below are organized in three sections defined by the following research sub-questions: (a) what are the relevant practices adopted by private and public organizations involved in commerce, science, and engineering for disaster recovery and long-term accessibility; (b) what best practices exist to manage data preservation; and (c) what efforts have been adopted by libraries for the purpose of archiving content to ensure (a) interoperability, (b) consistency, and the (c) safety and security of collections (Donaldson & Yakel, 2012, p. 55). All annotations presented portray the ideas and thoughts of the authors of the works listed.

What practices are used by private and public organizations for disaster recovery and long-term accessibility?

Chen, S.-S. (2001). The paradox of digital preservation. *Computer*, 24-28. doi: 10.1109/2.910890

Abstract. Preserving digital information is a problem plagued by short media life, obsolete hardware and software, slow read times of old media, and defunct Web sites. The paradox is: We want to maintain digital information intact, but we also want to be able to access this information in a dynamic use context. Chen explains that we lack

proven methods to ensure that the digital information will continue to exist, that we will be able to access this information using improved technology tools, or that accessible information is authentic and reliable.

The author asserts that failing to address the problems of preserving information in digital form is analogous to fostering cultural and intellectual poverty and squandering potential long-term gains that we should rightfully receive as a return on our professional, personal, and economic investments in information technology. Finding a solution to the tension between the creation context and the use context constitutes an important challenge.

Credibility. Su-Shing Chen is a professor of computer engineering and computer science at the University of Missouri-Columbia. His research interests include digital libraries, web technology, and bioinformatics. Chen received a PhD in mathematics from the University of Maryland at College Park (Chen, 2001, p. 6).

Computer is the flagship publication of the IEEE Computer Society which publishes highly acclaimed peer-reviewed articles written for and by professionals representing the full spectrum of computing technology from hardware to software and from current research to new applications (http://ieeexplore.ieee.org).

Summary. This article by Su-Shing Chen in the Computer journal, published by IEEE, discusses the challenges that face the preservation of digital information. Chen states that everyone has access to publishing content with the easy internet access many people enjoy, and accordingly many people and organizations invest time and effort in creating and capturing that content. He regrets that despite all this investment in technology, there is a glaring weakness in the information infrastructure, which is long-term preservation.

Chen presents examples of what is at stake including data lost throughout the last fifty years such as governmental information from the census bureau, NASA, presidential communications, and private companies. Chen compares traditional preservation to digital preservation, and explains the digital preservation paradox of wanting to maintain digital information intact, and at the same time wanting to access the information dynamically and with the most advanced tools. He lists the requirements and infrastructure for digital preservation including formats and styles, context, storage media, systems technology, the workflow process, and metadata policies. He concludes that both technology and expenditures will evolve, and failure to address the digital preservation problems is not an option. He asserts that we are compelled to meet the challenge to facilitate digital information preservation. Although this text was written in 2001, the assertions and context have not changed. This text remains a valuable source for the research into information preservation.

Gantz, J. (2008). The diverse and exploding digital universe. *IDC Whitepaper*. Retrieved from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Abstract. This white paper, sponsored by EMC, is an update of IDC's inaugural forecast of the digital universe published in March 2007. In this year's update we calibrate the size (bigger) and growth (faster) of the digital universe again, but we also explore some areas we only touched on last time. As before, we also seek to understand the implications for business, government, and society.

Credibility. John F. Gantz is the Senior Vice President of International Data Corporation (IDC), which is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology

markets. John Gantz is one of IDC's chief spokespersons on broad technology and market issues at major forums in the United States and around the world.

Prior to joining IDC in September 1992, Mr. Gantz was Vice President and Chief Analyst for Dataquest and director of that company's Software Research Group. Before joining Dataquest in 1991, Mr. Gantz was executive vice president of TFS, Inc., a custom research and consulting company that he co-founded in 1983. Mr. Gantz has been an IT industry analyst and columnist for more than 30 years. His publications and columns have appeared in Fortune, Forbes, Computerworld, Infoworld, Computer Graphics World, and Industry Week (Gantz, 2008).

Summary. In this text, Gantz provides an overview of the forecast of the digital universe and seeks to understand the implications for business, government, and society. Gantz starts by providing key findings from research including statistics about the digital world, from growth and diversity to digital information governance. In light of this explosion of the digital universe, Gantz presents IT organizations with three imperatives: (a) the need to transform existing relationships with business units to deal with information creation, storage, management, security, retention, and disposal; (b) the need to spearhead and develop policies for information governance; and (c) the need to rush new tools and standards to make information infrastructure as flexible, adaptable, and scalable as possible.

Gantz explores the growth of the information universe and investigates the physical repercussions of information overload with new devices and applications used to create content. He explores the evolution of storage media, diversity in digital content, and the dilemma facing enterprises of being responsible for content created by individuals.

Gantz estimates that 70% of digital content is created by individuals while the organizations are still held responsible for it. He cites real life examples of content posted on YouTube, peer-to-peer sharing sites, and virtual worlds. IDC estimates only about 35% of digital content is produced by enterprises, with much of that produced by workers on the road and away from the office. This leads to the dilemma of how to preserve and manage such content.

Gantz proposes lessons and methodologies for the enterprise, and ideas to plan for the future within a digital universe that is estimated to be ten times bigger than it was in 2008. He includes forecasting, estimating, and converting units to facilitate sizing the potential digital portfolio of an organization.

Gantz's work is important, as it is essential to understand the scope of the problem on hand when discussing information preservation. The first step to tackling the problem is identifying it.

Ludascher, B., Marciano, R., & Moore, R., A. (2001). Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *Sigmod Record, 30*(3), 54-63. doi:10.1145/603867.603876

Abstract. Digital archives are dedicated to the long-term preservation of electronic information and have the mandate to enable sustained access despite rapid technology changes. Persistent archives are confronted with heterogeneous data formats, helper applications, and platforms being used over the lifetime of the archive. This is not unlike the interoperability challenges, for which mediators are devised. To prevent technological obsolescence over time and across platforms, a migration approach for persistent at-chives is proposed based on an XML infrastructure. We extend current archival

approaches that build upon standardized data formats and simple metadata mechanisms for collection management, by involving high-level conceptual models and knowledge representations as an integral part of the archive and the ingestion/migration processes. Infrastructure independence is maximized by archiving generic, executable specifications of (i) archival constraints (i.e., "model validators"), and (ii) archival transformations that are part of the ingestion process. The proposed architecture facilitates construction of self-validating and self-instantiating knowledge-based archives. We illustrate our overall approach and report on first experiences using a sample collection from a collaboration with the National Archives and Records Administration (NARA).

Credibility. Professor Ludascher received his MS in Computer Science from the Technical University of Karlsruhe in 1992, and his PhD in Computer Science from the University of Freiburg in 1998, both in Germany. He currently teaches at the UC Davis Computer Science department. Professor Ludascher's primary research interests are in scientific data management, in particular scientific data integration, scientific workflow management, and knowledge-based (semantic) extensions thereof. He is an active contributor to several large scale research collaborations dealing with scientific data management, including the NSF/ITR Geosciences Network (GEON), the NSF/ITR Science Environment for Ecological Knowledge (SEEK), and the DOE Scientific Data Management Center. This research was sponsored by the National Science Foundation (SciDAC-SDM) (http://www.cs.ucdavis.edu).

Summary. In the paper, Professor Ludascher discusses the challenges that face persistent data archiving. He reviews the basics of archives, which include the mandate to preserve information so that it can be re-discovered, accessed, and presented at any time in the

future. Along with that comes the challenge of limited storage lifetime due to data decay. He then discusses the challenges associated with technological obsolescence of the infrastructure used to access the archived data.

Dr. Ludaescher, along with the National Archives and Records Administration and the San Diego Supercomputer Center, developed an information management architecture for digital archives to sustain access to data in the future and reinstate information in future platforms. Dr. Ludascher asserts that it is not enough to merely copy data at a bit level from one obsolete media to a newer one, but what is needed is to create recoverable archival representations that are infrastructure independent. Dr. Ludascher describes the architecture for infrastructure independent knowledge based archival and collection management to facilitate future accessibility.

Moore, R. (2008). Towards a theory of digital preservation. *The International Journal of Digital Curation*, 63-75. doi:10.2218/ijdc.v3i1.42

Abstract. A preservation environment manages communication from the past while communicating with the future. Information generated in the past is sent into the future by the current preservation environment. The proof that the preservation environment preserves authenticity and integrity while performing the communication constitutes a theory of digital preservation. We examine the representation information that is needed about the preservation environment for a theory of digital preservation. The representation information includes descriptions of the preservation management policies, the preservation processes, and the state information that is needed to verify the correct working behavior of the system. We demonstrate rule-based data grids that can

verify that prior policies correctly enforced preservation properties, while sending into the future descriptions of the current preservation management policies.

Credibility. Reagan W. Moore is director of the Data Intensive Cyber Environments Center (DICE Center) and a professor in the School of Library and Information Science at the University of North Carolina at Chapel Hill. He is a Chief Scientist for Data Intensive Cyber Environments at the Renaissance Computing Institute (RENCI), and President of the Data Intensive Cyberinfrastructure Foundation, which supports the open source community for the Integrated Rule-Oriented Data System (iRODS). An internationally recognized expert, he coordinates research efforts in development of data grids, digital libraries, and preservation environments. Current research activities include the use of data grid technology to automate execution of management policies and validate trustworthiness of repositories. His research is funded by the National Science Foundation under the Office of Cyberinfrastructure (OCI) program and the National Archives and Records Administration, under the Electronics Records Administration Research program, and other agencies. Previous positions include Associate Director for Data Intensive Computing, Director of the Knowledge and SRB Lab, and Manager of Production Systems at UC San Diego's Supercomputer Center, and computational plasma physicist at General Atomics (https://www.irods.org/index.php/Reagan_Moore).

The *International Journal of Digital Curation* (IJDC) is a peer-reviewed electronic journal entirely devoted to papers, articles and news items on the curation of digital objects and related issues (http://www.dcc.ac.uk/resources/curation-journals/ijdc).

Summary. Moore discusses the concept of preservation and describes it as communicating with the future. He recognizes that the future will bring newer, more

efficient, and more sophisticated technology than we have today, and accordingly, preservation entails ensuring that the future technology can adequately read and access today's data. Moore advises that the preservation environment will need to incorporate new types of storage systems, protocols for accessing data, new encoding formats, and new standards. Accordingly, the challenge we are faced with today is incorporating new technology effectively while maintain the properties that guarantee preservation, such as authenticity, integrity, and chain of custody.

Moore views preservation as communication from the past, which requires information professionals to make assumptions about prior applications and preservation policies since any claims of data integrity rely completely on prior actions taken in the past. Through ensuring past preservation actions, a theory of preservation becomes possible based on the definition of the minimal set of processes needed to implement preservation policies. Moore explores the concepts of infrastructure independence to ensure data accessibility in the future regardless of platform and details the assessment criteria needed by the preservation community to validate repositories. Moore's platform independence is a key to future accessibility and is essential to the understanding of information preservation.

Moore, R., Rajasekar, A., & Wan, M. (2005). Data grids, digital libraries, and persistent archives: An integrated approach to sharing, publishing, and archiving data. *Proceedings of the IEEE, 93*(3), 578-588. doi:10.1109/JPROC.2004.842761

Abstract. The integration of grid, data grid, digital library, and preservation technology has resulted in software infrastructure that is uniquely suited to the generation and management of data. Grids provide support for the organization, management, and

application of processes. Data grids manage the resulting digital entities. Digital libraries provide support for the management of information associated with the digital entities. Persistent archives provide long-term preservation. We examine the synergies between these data management systems and the future evolution that is required for the generation and management of information.

Credibility. Reagan W. Moore received his BS degree in physics from the California Institute of Technology, Pasadena, in 1967 and the PhD degree in plasma physics from the University of California, San Diego, in 1978. He is Director for Data Intensive Computing Environments at the San Diego Supercomputer Center (SDSC), University of California, San Diego. He coordinates research efforts on digital libraries, data grids, and persistent archives.

Notable collaborations include the National Science Foundation (NSF), National Virtual Observatory, the NSF National Science Digital Library persistent archive, the NSF Southern California Earthquake Center community digital library, the Department of Energy (DOE) Particle Physics Data Grid, the NHPRC Persistent Archive Testbed, and the NARA Prototype Persistent Archive (Moore, Rajasekar, & Wan, 2005, p. 587).

IEEE is the world's largest professional association dedicated to advancing technological innovation and excellence for the benefit of humanity. IEEE and its members inspire a global community through IEEE's highly cited publications, conferences, technology standards, and professional and educational activities.

(http://www.ieee.org/about/index.html)

Summary. In this IEEE paper, Moore, Rajasekar, and Wan describe the integration of grid, grid data, digital libraries, and preservation technology and their role in software

infrastructure that is well suited for managing data. They examine the interaction between data management systems and future evolution that is required for the generation and management of information, and introduce data grids which support massive data collections distributed across multiple organizations and based upon generic data management infrastructures. Precise definitions are provided of data, information, and knowledge, in order to distinguish among them for the purpose of preservation and accessibility. They discuss assigning semantic labels to digital entities to assert properties, which are then assigned to a database to provide context for interpreting them. One major research issue they focus on is the integration of knowledge management systems with existing data and information management systems as well as integrating digital libraries with data grids. The authors go to great lengths to discuss these ideas and the implementation on a large scale. Moore, Rajasekar, and Wan conclude that the integration of data grids, digital libraries, and persistent archives is forcing the continual evolution of grid technology and that the ability to manage them will require further evolution of the grid technology.

Rabinovici, S., Baker, M., Cummings, R., Fineberg, S., & Marberg, J. (2011). Towards SIRF: Self-contained information retention format. *Proceedings of the 4th Annual International Conference on Systems and Storage*. New York: ACM. doi: 10.1145/1987816.1987836

Abstract. Many organizations are now required to preserve and maintain access to large volumes of digital content for dozens of years. There is a need for preservation systems and processes to support such long-term retention requirements and enable the usability of those digital objects in the distant future, regardless of changes in technologies and designated communities. A key component in such preservation systems is the storage

subsystem where the digital objects are located for most of their lifecycle. We describe SIRF (Self-contained Information Retention Format) -- a logical storage container format specialized for long term retention. SIRF includes a set of digital preservation objects and a catalog with metadata related to the entire contents of the container as well as to the individual objects and their interrelationship. SIRF is being developed by the Storage Networking Industry Association (SNIA) with the intention of creating a standardized vendor-neutral storage format that will be interpretable by future preservation systems and that will simplify and reduce the costs of digital preservation.

Credibility. Simona Rabinovici-Cohen is a research staff member at the IBM Research Laboratory in Haifa, Israel. She holds MSc and BSc degrees in computer science, both from the Technion - Israel Institute of Technology. Rabinovici-Cohen leads Preservation DataStores in the Cloud that provides preservation-aware cloud based storage services to ENSURE EU integrated project. In standardization efforts, Rabinovici-Cohen co-chairs Long Term Retention technical working group in SNIA that develops the Self-contained Information Retention Format (SIRF) (http://researcher.watson.ibm.com/researcher/view.php?person=il-SIMONA).

ACM is the world's largest educational and scientific computing society, delivering resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources (ww.acm.org).

Summary. In this paper, Rabinovici-Cohen and her colleagues who represent a number of firms from a cross-section of the technology industry, discuss preservation efforts and

the need for them. They introduce Self-contained Information Retention Format (SIRF) a logical storage container that is intended for long-term retention.

They discuss the requirement by a growing number of large organizations to preserve large volumes of digital content for an extended amount of time, while maintaining access to that information for regulatory and compliance purposes. They define long-term retention as the ability to sustain the accessibility, understandability, and usability of digital objects in the distant future regardless of changes to technologies or to the user communities. They divide the preservation challenge into logical preservation and bit preservation. Bit preservation is defined as the ability to retrieve data from the physical storage media despite any damage or corruption to the media. Logical preservation is preserving the understandability and usability of the data despite unforeseeable changes that will take place in technology or users in the future.

Rabinovici-Cohen et al. introduce SIRF as a technology similar to physical archiving where documents are gathered and organized in some manner, placed in a container or folder, which is then marked and placed in a known location, and information about it is placed in a finding aid such as a catalog. SIRF is the digital equivalent to the storage container that defines a series, which can be labeled with standard information in a defined format to allow retrieval as needed.

Rabinovici-Cohen et al. present some use cases and requirements for SIRF, as well as the different actors who will be involved in the process. SIRF is one of many powerful tools that can make future accessibility possible.

What are best practices designed to ensure data is usable for as long as it needs to be kept within the organization (i.e., data preservation)?

Berman, F. (2008). Got data?: A guide to data preservation in the information age. *Communications of the ACM* 51, 50-56. doi:10.1145/1409360.1409376

Abstract. Tools for surviving a data deluge to ensure your data will be there when you need it.

Credibility. Francine Berman is the director of the San Diego Supercomputer Center, professor of Computer Science and Engineering, and High Performance Computing Endowed Chair in the Jacobs School of Engineering at the University of California, San Diego. She is also co-chair of the Blue Ribbon task force on Sustainable Digital preservation and access (Berman, 2008, p. 56), which is supported by the National Science Foundation, Library of Congress, and other foundations.

ACM is the world's largest educational and scientific computing society, delivering resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources (ww.acm.org).

Summary. In this paper, published by the Communications of the ACM, Berman presents the reader with the concept of the proliferation of digital data and tools for surviving it to ensure data is there when it is needed. She asserts that data is fragile, as evidenced by crashing drives, obsolete media, loss, damage, and unavailability. She emphasizes the importance of securing data and provides examples from the news that

further highlight dependence on electronic information. Berman then explores key trends and issues associated with preserving digital data.

Berman discusses digital infrastructure, including distributed computer, information, and communication technology. She states that users want to store and use their data for periods spanning the short-term to the long-term, and they want it to be available to their collaborators and communities. She describes trends in data production and creation, citing research that estimates that the amount of data created will surpass Avogadro's number by 2023. Accordingly, storing, accessing, managing, preserving, and dealing with digital data is a fundamental need requiring attention. Berman predicts four significant trends that reflect the larger environment in which data cyber-infrastructure is evolving: (a) more digital data is being created than there is storage, (b) more policies and regulations require data preservation, (c) storage costs are decreasing, and (d) the increasing commercialization of digital data storage and services.

Accordingly, she asserts that the importance of *appraisal* of data to select what needs to be preserved will be critical to communities in research and education, however, the decreasing cost of storage needs to be balanced with curation, annotation, and professional expertise.

Donaldson D., & Yakel, E. (2012). Secondary adoption of technology standards: The case of PREMIS. *Archival Science, 13*(1). doi: 10.1007/s10502-012-9179-0

Abstract. While archival scholars have identified some of the most important steps for deciding to use and implement metadata standards in archives, very little systematic empirical investigation within the archival science literature regards either how implementation processes actually unfold or the factors affecting implementation.

This article analyzes the organizational factors and processes that come into play during implementation of metadata standards, using PREservation metadata: implementation strategies (PREMIS) as an exemplar. Adapting a theoretical framework for secondary adoption of technologies from Gallivan (Database Adv Inf Syst 32(3):51, 2001), the authors apply their model to the PREMIS technology standard and investigate PREMIS implementation by projects/programs on the Library of Congress PREMIS Implementation Registry. Using data from a series of in-depth semi-structured interviews, the authors develop a model for the secondary adoption of PREMIS and outline implications for the secondary adoption of technology standards based on the results of this study.

Credibility. Devan Ray Donaldson is a doctoral candidate in the School of Information at the University of Michigan, Ann Arbor. He conducts research in the areas of digital preservation and curation focusing on preservation management, preservation metadata, preservation repositories and users. Donaldson has been a Bill and Melinda Gates Millennium Scholar since 2002, a Horace H. Rackham Merit Fellow since 2008 and an Edward Alexander Bouchet Graduate Honor Society Member since 2012. He earned a MS in Library Science from the University of North Carolina and a BA in History from the College of William and Mary in Virginia (Donaldson & Yakel, 2012, p. 83). Elizabeth Yakel is Associate Professor in the School of Information at the University of Michigan.

The *Archival Science* journal covers all aspects of archival science theory, methodology, and practice. Moreover, it investigates different cultures and promotes the exchange and comparison of concepts, views, and attitudes around the world. Its scope encompasses the

entire field of recorded process-related information, analyzed in terms of form, structure, and context. To meet its objectives, the journal draws from scientific disciplines that deal with the function of records and the way they are created, preserved, and retrieved; the context in which information is generated, managed, and used; and the social and cultural environment of records creation at different times and places.

Summary. In this paper, Donaldson and Yakel discuss metadata standards for archives. They state the importance of standards to reflect the most current knowledge about professional practices and increasing interoperability, consistency, and safety and security of collections. They regret that despite the importance of standards, they can easily be ignored or diffused due to broad changes that adopting standards entails. Increasingly, interest in preservation metadata has been seen due to concerns over the long-term maintenance of digital objects.

Donaldson and Yakel identify the most important steps for deciding to use and implement metadata standards in archives. The first step is to consider who developed the metadata standard, why it was developed, to which entities it relates, and what functions it supports. Secondly, an institution should consider its organizational context and how the standard fits into that context and adapt the standard accordingly. Additionally, institutions must create application profiles, which formally declare how they intend to implement the metadata standards.

Donaldson and Yakel agree that none of these activities are trivial and without challenges, particularly for institutions wishing to make the most of metadata standards to manage digital objects while increasing interoperability and consistency among

repositories. Challenges identified by scholars in implementing metadata are acknowledged, and comparisons to other standards in use today are made.

The PREMIS standard is then introduced, including background, methodology, findings, and implementation. In conclusion, they propose refinements to some existing standards to provide a more realistic framework that can explain the interplay among organizational context variables, attributes of managers' implementation strategies, and other characteristics. They also propose recommendations for managers and implementers.

Gladney, H. (2006). Principles for digital preservation. *Communications of the ACM*, *49*(2), 111-116. doi: 10.1145/1113034.1113038

Abstract. The immense investments in creating and disseminating digitally represented information have not been accompanied by commensurate effort to ensure the longevity of information of permanent interest. Asserted difficulties with long-term digital preservation prove to be largely underestimation of what technology can provide. We show how to clarify prominent misunderstandings and sketch a *Trustworthy Digital Object (TDO)* method that solves all the published technical challenge.

Credibility. H. M. Gladney is the president of HMG Consulting in Saratoga, CA, and the publisher of Digital Document Quarterly and former staff members of IBM Almaden Research Center, San Jose, CA. His consulting specialties include enterprise digital document management, especially long-term digital preservation, document security, authenticity, trust, and intellectual property law and management (http://hgladney.com/). *ACM* is the world's largest educational and scientific computing society, delivering resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing

profession with leading-edge publications, conferences, and career resources (ww.acm.org).

Summary. In this peer-reviewed paper, Gladney explains some principles for data preservation. He starts by reviewing the current status of information, asserting that most information today is "born digital", but despite that, little is provided to ensure its long term integrity. He states that there needs to be an urgency for preserving authentic digital works. He stresses the need to include the information technology community in the preservation efforts, and not limit it to scholars and artists.

The focus in this paper is on challenges created by technological obsolescence and the demise of information providers. Gladney lists challenges exposed by prior work, then distinguishes between original and authentic work. He describes the Trusted Digital Object (TDO) methodology which focuses on preserving objects, defines methods for making their authenticity reliably testable, and assures that eventual users will be able to render or otherwise use their contents.

The paper addresses what he thinks is missing from the U.S. digital preservation plan and concludes with the assertion that TDO focuses on end users' needs by encapsulating schemes for digital preservation objects. This allows preservation of any type of digital information and is as efficient as any competing solution.

Goth, G. (2012). Preserving digital data. *Communications of the ACM, 55*(4), 11-13. doi:10.1145/2133806.2133811

Abstract. The article looks at the issue of digital data preservation, with particular focus on the rendering, curating, and archiving of scientific research. It is argued that funding is needed for a data preservation infrastructure for storing and maintaining future

accessibility of archived information. Several digital data measures are discussed, including the International Organization for Standardization (ISO) proposal for a Digital Preservation Interoperability Framework specification, the U.S. White House of Science and Technology Policy Request for Information about public digital data access, and the Integrated Rule-Oriented Data System (iRODS) developed by Data Intensive Cyber Environments (DICE) researchers at the University of North Carolina and the University of California, San Diego.

Credibility. Gregory Goth is an Oakville, CT-based writer who specializes in science and technology, he was the staff news editor at IEEE Computer Society.

ACM is the world's largest educational and scientific computing society, delivering resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources (ww.acm.org).

Summary. In this paper published in the Communications of the ACM, Goth analyzes data growth in light of the need for more policy coordination. He seeks a way to make artifacts useful across space and time. In the past, archiving was simple, it only required preserving static artifacts such as printed pages and photographs. Nowadays, artifacts are more complex in nature, they require a computer to render. The sheer number of agencies and efforts devoted to tackling the issue of preserving digital content highlight the size of the task at hand. Goth attests that the irony facing these efforts is that we are so focused on the next new discovery, we lose sight of and ignore old data. Goth cites researchers stating that the biggest challenge facing data preservationists is the

infrastructure to make it all happen. He states that agencies that provide funding generally tend to provide funding for short-term research, but almost never for preservation.

Goth is encouraged by the challenges to the existing cultural models by some more comprehensive national initiatives and he lists a few examples of such initiatives by governments and agencies. He lists some examples of standards and research aimed at digital preservation. Goth's concern over lack of funding and attention towards digital preservation of the past is the essence of the problem addressed by this research paper. Identifying the problem is the first step to finding a resolution.

Hart, P. & Liu, Z. (2003). Trust in the preservation of digital information. *Communications of the ACM 46*, 6 93-97. doi:10.1145/777313.777319

Abstract. We live in a digital world. With an increasing amount of information being created, stored, and distributed in digital formats, preservation of digital information is a central concern. A recent wave of literature addressing the subject has focused on the fragility of digital media, technological obsolescence, and standards, but little attention has been given to the most critical barrier in the preservation of digital information: the potential conflicts between the new reality of digital information and the expectations of people.

We address this issue here, describing the results of our survey of individuals with intensive experience in handling digital information and applying a concept derived from our analysis of monetary currency, called "institutional guarantee," to the development of trusted systems for the preservation of digital information.

Credibility. Peter E. Hart is chairman and president of Ricoh Innovations, Inc., Menlo Park, CA, and Senior Vice President of Ricoh Company, Ltd., Tokyo, Japan.

Ziming Liu is an assistant professor at the School of Library and Information Science, San Jose State University, San Jose, CA (Hart & Liu, 2003, p. 97).

ACM is the world's largest educational and scientific computing society, delivering resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources (ww.acm.org).

Summary. In this Communication of the ACM, Hart and Liu address the problem of gaining people's trust in the challenge of digital preservation. Hart and Liu perform a study of 110 individuals that exposes the fact that despite the advances in digital technology, the majority of respondents still keep paper copies. Participants in the survey listed five factors supporting their lack of trust in digital preservation: (a) inaccessibility, (b) lack of tangibility, (c) fluidity, (d) short preservation periods, and (e) privacy and security.

Hart and Liu stress that the creation of an institutional guarantee for trusted digital preservation is essential to increasing people's confidence in digital media. They agree that trust, as important as it is, is fragile, and people are unwilling to take risks, which leads to a vicious cycle of lack of confidence and the need for trust.

The authors list four characteristics of trust, including (a) it increases with familiarity, (b) it is linked to a given condition, (c) it requires accountability and tangibility, and (d) it is often associated with scale.

The authors then explain some generic trust requirements needed for digital preservation, and conclude that although the cost of electronically storing documents is dipping below the cost of storing paper documents, the cost and convenience advantages of digital archiving require solutions to an issue as old as humanity and as new as the next "next thing": trust. This paper fills in a gap in many of the other documents in that it addresses the human psychological aspect of data preservation, not just the technical side.

Kuny, T. (1997). A digital dark ages? Challenges in the preservation of electronic information. *63RD IFLA Council and General Conference*. Retrieved from http://archive.ifla.org/IV/ifla63/63kuny1.pdf

Abstract. Digital collections facilitate access, but do not facilitate preservation. Being digital means being ephemeral. Digital places greater emphasis on the here-and-now rather than the long-term, just-in-time information rather than just-in-case. The research program for digital preservation has only recently been initiated to develop strategies, guidelines, and standards. Although tremendous work has been undertaken in defining the problems and challenges, much more remains to be done, and the tough task of actually doing digital preservation (and digital rescue) remains ahead. A critical appraisal of where we are vis-a-vis our digital culture, and what we want for the future something which may not be defined in technical terms at all is required both inside and outside of the library and archival professions. If history and cultural heritage are to be important, then it will likely fall to librarians and archivists, the monastic orders of the future, to ensure that something of the heady days of our "digital revolution" remains for future generations. The challenges to digital preservation are considerable and will require a

concerted effort on the part of librarians and archivists to rise up to these challenges and assert in public forums the importance of protecting a fragile digital heritage.

Credibility. *The International Federation of Library Associations and Institutions* (IFLA) is the leading international body representing the interests of library and information services and their users. It is the global voice of the library and information profession. Founded in Edinburgh, Scotland, in 1927 at an international conference, IFLA now has over 1500 members in approximately 150 countries around the world. IFLA aims to promote high standards of provision and delivery of library and information services, encourage widespread understanding of the value of good library & information services, and represent the interests of members throughout the world.

Summary. Kuny opens his report by citing George Orwell's 1949 novel, 1984: *Who controls the past controls the future. Who controls the present controls the past,* which is a great introduction to the topic he discusses. Kuny starts by reviewing the history of preservation dating back to monks and monasteries in the middle ages and their vital role in preserving knowledge. Kuny disregards the claim that the internet changes everything, instead he calls on the readers to remember the past by preserving historic records in the electronic era. Kuny likens our current state of affairs with that faced by barbarians at the gate of something not fully understood. Kuny provides numerous observations to back up his claim that we're living in the digital dark ages including: (a) an enormous amount of digital data has already been lost, (b) waves of new digital information will be emerging as baby boomers retire and start writing their life experiences, (c) technology is essentially obsolete every 18 months, and (d) there is an explosion in document and media formats.

Kuny claims that although we do not understand how to archive digital data and sustainable solutions to preservation are not available, that preservation of digital materials is not complex as long as the relationship between hardware, software, and people is maintained. The problem is with forces that pull each of these elements away from each other, such as when software and hardware become outdated, migrating information may require expensive recoding, and organizations lack resources and motivation to address the problems. This creates a situation where the object is left in digital space, trapped in an obsolete format, captured on an unreadable medium, or lacking the administrative capacity, resources, or willingness to refresh the data. Kuny agrees with others that digital objects require frequent refreshing and recopying to new storage media. Keeping the original digital artifact and moving it from one storage medium to another is important but it is essential to also translate it into new formats or structures.

Kuny lists some recommendations to libraries on what needs to be done to address this problem, including digital triage, rescue operations, legal rights management, and working together. He concludes with another quote, this one from George Santayana's *The Life of Reason,* that those who can't remember the past are condemned to repeat it. Although relatively old, Kuny's paper provides insight into the importance and need for digital preservation to ensure our cultural heritage is not lost.

Kunze, J. (2004). Ark. *Computers in libraries, 24*(2), 18. Retrieved from http://search.ebscohost.,com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=12181345&site=ehost-live&scope=site

Abstract. This article presents information on Archival Resource Key (ARK) specification, a naming scheme designed to facilitate the high-quality and persistent identification of information objects. A founding principle of the ARK is that persistence is a matter of service and is neither inherent in an object nor conferred on it by a particular naming syntax. An ARK is a special kind of uniform resource locator that connects users to three things: the named object, its metadata and the provider's promise about its persistence. The ARK scheme recognizes that two important classes of name authority affect persistence: original assigners of names and current providers of mapping services. Work on ARK started at the U.S. National Library of Medicine, Lister Hill Center, Computer Science Branch. The ARK specification is stable as of February 2004, but subject to ongoing refinement and extension. The disadvantage of ATK is that tool support is immature. Its advantages include: low buy-in cost, available connection not only to objects, but also to their providers' metadata and commitment statements and the core identity of ARK can be recovered by stripping off the hostname.

Credibility. John Kunze is the associate director of the California Digital Library (CDL) and is a preservation technologist. With a background in computer science and mathematics, he wrote software that comes pre-installed on Linux and Apple operating systems. He has also contributed heavily to the standardization of URLs, Dublin Core metadata, and web archiving. Kunze provides analysis, design, programming, leadership,

and communications support for all aspects of CDL's digital preservation and curation activities (http://www.cdlib.org/contact/staff_directory/jkunze.html).

Computers in libraries (CIL) is a monthly magazine that provides complete coverage of the news and issues in the rapidly evolving field of library information technology. Focusing on the practical application of technology in community, school, academic, and special libraries, CIL includes discussions of the impact of emerging computer technologies on library systems and services, and on the library community itself (http://www.infotoday.com/cilmag/).

Summary. In this article published in the monthly trade journal, Kunze introduces Archival Resource Keys (ARK), a naming scheme designed to facilitate high quality and persistent identification of information objects. An ARK is a special kind of uniform resource locator (URL) that points users to three things, a named object, its metadata, and the provider's promise of persistence. Kunze explains the difference between ARK and other schemes such as URN, DOI, and PURL, and provides examples of how it is used and formatted. Kunze then explains the benefits of the ARK system in light of longevity and accessibility. He concludes by crediting the team and organizations that sponsored the research. ARK is a promising scheme that can help data archiving to assist in retrieval and preservation.

Lorie, R. A. (2001). Long term preservation of digital information. *1st ACM/IEEE-CS joint Conference on Digital libraries*, 346-352. doi:10.1145/379437.379726

Abstract. The preservation of digital data for the long term presents a variety of challenges from technical to social and organizational. The technical challenge is to ensure that the information, generated today, can survive long term changes in storage

media, devices and data formats. This paper presents a novel approach to the problem. It distinguishes between archiving of data files and archiving of programs (so that their behavior may be reenacted in the future).For the archiving of a data file, the proposal consists of specifying the processing that needs to be performed on the data (as physically stored) in order to return the information to a future client (according to a logical view of the data). The process specification and the logical view definition are archived with the data. For the archiving of a program behavior, the proposal consists of saving the original executable object code together with the specification of the processing that needs to be performed for each machine instruction of the original computer (emulation).In both cases, the processing specification is based on a Universal Virtual Computer that is general, yet basic enough as to remain relevant in the future.

Credibility. The *Association for Computing Machinery* (ACM) is the world's largest educational and scientific computing society, delivering resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources (ww.acm.org). Raymond Lorie is a researcher at IBM, Almaden.

Summary. Raymond Lorie discusses some challenges to long-term preservation of digital information. With the change from paper as the main medium to digital media, the challenge has been to ensure that the information can survive long-term changes in storage media, devices, and data formats. Lorie presents three conditions to ensure successful recovery in the future: (a) the file must be physically intact, (b) a device must

be available to read the bit stream, and (c) the bit stream must be correctly interpreted. Lorie then introduces different types of documents and several cases to consider. Lorie covers some previously proposed schemes such as conversion and emulation, and then presents his proposal, which is to differentiate between data archival and program archival. The paper then goes into depth about the method proposed by Lorie for data archival.

Maniatis, P., Roussopoulos, M., Giuli, T., Rosenthal, D. & Baker, M. (2005). The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on Computer Systems*, *23*(1), 2-50. doi: 10.1145/1047915.1047917

Abstract. The LOCKSS project has developed and deployed in a world-wide test a peer-to-peer system for preserving access to journals and other archival information published on the Web. It consists of a large number of independent, low-cost, persistent Web caches that cooperate to detect and repair damage to their content by voting in *opinion polls*. Based on this experience, we present a design for and simulations of a novel protocol for voting in systems of this kind. It incorporates rate limitation and intrusion detection to ensure that even some very powerful adversaries attacking over many years have only a small probability of causing irrecoverable damage before being detected.

Credibility. Petros Maniatis is a Senior Research Scientist at Intel Labs. He is currently a member of the Intel Science and Technology Center on Secure Computing at UC Berkeley. Between 2003 and 2011, he was a research scientist at Intel Labs Berkeley. He received his MSc and PhD from the Computer Science Department at Stanford University. Prior to Stanford, he obtained his BSc with honors at the Department of

Informatics of the University of Athens in Greece. His research interests lie primarily in the confluence of distributed systems, security, and fault tolerance.

ACM is the world's largest educational and scientific computing society, delivering resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources (ww.acm.org).

Summary. In this technical paper, Maniatis et al. discuss the migration of academic publishing to the web which is forcing libraries to transition from purchasing copies, to renting access to the publishers' copies. This leads to the dilemma of the lack of guarantee of long-term access to the material. After millennia of experience with physical documents, librarians consider it one of their responsibilities to ensure future long-term access to content. Maniatis et al. propose the lots of copies keep stuff safe (LOCKSS) system, which models the physical document system and applies it to web-published academic journals, providing tools for libraries to take custody of the material to which they subscribe, and to cooperate with other libraries to preserve it and provide access.

In this paper, Maniatis et al. present a design and simulations for a new peer-to-peer opinion poll protocol that addresses scaling and attack resistance issues. The proposed protocol is based on the authors' experience with the deployed LOCKSS system and the special characteristics of such a long-term large-scale application. Their system with its time horizon of many decades and lack of central control, presents both unusual requirements, such as the need to avoid long-term secrets like encryption keys, and

unusual opportunities, such as the option to make some system operations inherently very time-consuming without sacrificing usability. The system resists both random failures, and deliberate attacks over a long time. LOCKSS and its underlying principles are novel and will prove useful in the design of other long-term large-scale applications operating in hostile environments, which is an important facet to consider when dealing with long-term document preservation.

McClure, M. (2006). Archive-it 2. *EContent*, *29*(8), 14-15. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=22452263&site=ehost-live&scope=site

Abstract. The article analyzes the problem of preserving web content. Collecting URLs is not enough to keep track on valuable content. To make digital content preservation possible, Internet Archive, a nonprofit company has led a charge to effectively capture and store web content. Internet Archive has focused on ensuring the availability and accessibility of Internet content by creating an Internet library to store digital content. It launched a service called Archive-It to help organizations seek a way to archive web content. Internet Archive's complete library contains 65 billion pages of web content including books, moving images, and software.

Credibility. Marji McClure is a freelance writer based in Connecticut. *EContent* is a leading authority on the businesses of digital publishing, media, and marketing, targeting executives and decision-makers in these fast-changing markets. By covering the latest tools, strategies, and thought-leaders in the digital content ecosystem, EContent magazine and EContentmag.com keep professionals ahead of the

curve in order to maximize their investment in digital content strategies while building sustainable, profitable business models (http://www.econtentmag.com/About_Us).

Summary. McClure investigates the daunting task of preserving web content in light of constant changes. Simply collecting uniform resource locators (URLs) is not enough to keep up with the valuable content. McClure introduces an archiving tool developed by Internet Archive that aims to ensure the availability and accessibility of internet content by creating an internet library to permanently store digital content for anyone to view at any time. McClure addresses the *fallacy* that if something is on the web, it will stay there, and stresses the need for people to understand that the web reflects who we are, accordingly, we don't want to lose any of it.

McClure estimates the library to currently hold sixty five billion pages including books, moving images, and software; the documents are stored in repositories around the world. Archive-It was designed mainly for institutions that have a mandate to archive their web content and that lack the resources to do so, however, they are collaborating with institutions to save material that they normally wouldn't save on their own. Archive-It aims to make sure that all of this knowledge is not lost.

Internet Archive has plans to expand the reach of the Archive-It service by targeting smaller entities, such as independent researchers, local libraries, and small non-governmental organizations to disseminate and access information embodied in its Archive-It service that anybody or any organization can use.

McGath, G. (2013). The format registry problem. *Code4Lib Journal*, 1-10. Retrieved from http://journal.code4lib.org/articles/8029

Abstract. File format identification is an important issue in digital preservation. Several noteworthy attempts, including PRONOM, GDFR, and UDFR, have been made at creating a comprehensive repository of format information. The sheer amount of information to cover and the constant introduction of new formats and format versions has limited their success. Alternative approaches, such as Linked Data and offering limited per format information with identifiers that can be used elsewhere, may lead to greater success.

Credibility. Gary McGath is a software engineer for the Harvard Library. The validation and analysis tool JHOVE, which is widely known in the library and preservation communities, is mostly his code. Currently, he is an independent software developer with strong connections in the digital preservation world. He recently published a book titled *Files that Last: Digital preservation for every geek* on digital preservation for home users, organizations, and administrators, which contains information on backup, archiving, format choice, organization, and more! (http://www.kickstarter.com/profile/garymcgath).

The *Code4Lib Journal* exists to foster community and share information among those interested in the intersection of libraries, technology, and the future (http://journal.code4lib.org/mission).

Summary. In this technical paper by Gary McGath, published by the Code4Lib Journal, the author discusses the file format problem and its effect on digital content. McGath refers to the thousands of formats with tens of thousands sub-varieties. This makes the

life of the digital preservationist difficult as they cannot tell whether a valid file is one which can easily be opened and will stay that way in the reasonable future, or are special measures needed to ensure its readability. McGath starts with the early days of computing where each system had its own format, and files couldn't be copied from one system to another due to incompatibility. The only way to preserve digital content in those days was to print it off on paper. As the number of incompatible operating systems declined, long-term retention of files became a possibility, and libraries and archives have been particularly aware of the need to keep information about formats after the software which creates them goes out of use.

McGath presents some examples of attempts at establishing repositories of file format specifications, documentation, and related software for format-specific materials related to migration as a preservation strategy. McGath presents information about PRONOM, GDFR, UDFR, Wikipedia, Just Solve the Problem, and Unification by Linked Data. He concludes that the format registry problem is an ongoing one where new formats develop and old ones change both officially and unofficially. Accordingly, no institution has enough resources capable of gathering a complete registry of formats and information. No one is going to *just solve the problem* once and for all. Ongoing efforts by many different people working independently will be necessary to keep up with the growing variety of formats.

National Science Board (2005). Long-lived digital data collections: Enabling research and education in the 21st century. *National science foundation*. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

Abstract. The primary purpose of this report is to frame the issues and to begin a broad discourse. Specifically, the NSB and NSF working together – with each fulfilling its respective responsibilities – need to take stock of the current NSF policies that lead to foundation funding of a large number of data collections with an indeterminate lifetime and to ask what deliberate strategies will best serve the multiple research and education communities. The analysis of policy issues in Chapter Four and the specific recommendations in Chapter Five of this report provide a framework within which that shared goal can be pursued over the coming months. The broader discourse would be better served by interaction, cooperation, and coordination among the relevant agencies and communities at the national and international levels. Chapters Two and Three of this report, describing the fundamental elements of data collections and curation, provide a useful reference upon which interagency and international discussions can be undertaken. The Board recommends that the Office of Science and Technology Policy (OSTP) take the lead in initiating and coordinating these interagency and international discussions.

Credibility. The *National Science Board* (NSB) is the governing board of the National Science Foundation (NSF) and policy advisors to the president and congress. The NSF is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; and to secure the national defense." The NSF is the funding source for approximately 20 percent of all federally supported basic research conducted by America's colleges and universities. In many fields such as mathematics, computer science and the social sciences, NSF is the major source of federal backing (http://www.nsf.gov/about/glance.jsp).

Summary. In this paper published by the NSB, the researchers discuss the fundamental changes brought on by the advancements in information technology, specifically, data collections, which have enabled the analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. This report focuses on long-lived data collections that meet the following definitions: (a) *data* is used to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc.; (b) *collection* is used to refer not only to stored data but also to the infrastructure, organizations, and individuals necessary to preserve access to the data; (c) *digital collections* that are the focus for this report are limited to those that can be accessed electronically; and (d) *long-lived* refers to a period of time long enough for there to be concern about the impacts of changing technology.

The NSB identifies the growing importance of such collections for research and education, and their potential for broadening participation in research at all levels. This report frames the issues and begins the discussion into the matter. The report is divided into five chapters: (a) introduction, (b) the elements of the digital data collection universe, (c) roles and responsibilities of individuals and institutions, (d) perspectives on digital data collections policy, and (e) findings and recommendations. The report concludes with a set of recommendations, including steps recommended by the NSF such as requiring research proposals and activities to state intentions so that peer reviewers can evaluate the data management plan, ensuring that education and training in the use of digital collections is available, and developing and growing the career path for data scientists to ensure that the research enterprise includes a sufficient number of high-

quality data scientists. The role of governmental organizations with authority such as the NSF and NSB is vital to ensure data preservation is a successful endeavor.

Ray, J. (2012). The rise of digital curation and cyber infrastructure: From experimentation to implementation and maybe integration. *Library Hi Tech, 30*(4), 604-622. doi: 10.1108/07378831211285086

Abstract. The terms *digital curation* and *cyberinfrastructure* have been coined in the last decade to describe distinct but related concepts of how data can be managed, preserved, manipulated and made available for long-term use. This paper aims to examine these. Design/methodology/approach – The paper considers the origins of both terms and the communities that have been engaged with each of them, traces the development of the present digital environment in the USA and considers what this may mean for the future. Findings – The paper reveals that each term has important attributes that contribute to a comprehensive understanding of the digital knowledge universe. Originality/value – The paper reveals information about the development of digital preservation.

Credibility. Joyce Ray was Associate Deputy Director of Library Services at the US Institute of Museum and Library Services (IMLS) from 1997 to 2011. She is currently on the museum studies faculty of Johns Hopkins University. She is a certified archivist and worked at the US National Archives and Records Administration for ten years before joining IMLS. She has a PhD in History and a Master's degree in Library and Information Science, both from the University of Texas at Austin (Ray, 2012, p. 622). The *Library Hi Tech* journal focuses upon computing and technology for the library community. It is international in scope and defines technology in the broadest possible terms to include the full range of tools employed by librarians and their customers. The

majority of journal issues are themed, thus allowing for extensive in-depth coverage and analysis of key areas (http://www.emeraldinsight.com/products/journals/journals.htm?id=lht&PHPSESSID=fpqc0ld07sqnqbjsjli32i91n3).

Summary. In this paper published in the Library Hi Tech journal, Ray discusses digital curation in libraries from the perspective of an archivist who has been involved with the development of digital libraries and repositories over the period of fifteen years.

Ray admits that the term *digital curation* has been somewhat controversial with some opposition coming from fields that had already employed curators in job titles, such as museums. Despite that, the use of the terms *content curation* and *web curation* to describe the selection and posting of digital content on fashion blogs and social media sites seems to have overshadowed objections to the use of the term by other professions.

Ray proceeds to compare the definition of *digital curation* with that of *cyberinfrastructure*, and then investigates the origins, curation lifecycles, repositories, the OAIS reference model, and federal funding. Ray examines institutional repositories and the rise of digital curation as a profession including the recommendation that US federal agencies promote a data management planning process for projects that generate preservation data with backing from the National Science Foundation (NSF).

Ray transitions into discussing digital curation tools including data management plans, digital curation profiles, variable media questionnaire, and data visualization tools. She concludes her paper by stating her intentions for the paper, which were to scan the last decade and a half of the digital environment to show how its development has influenced

the present and to make explicit the connection between today's cyberinfrastructure and the principles and practices of digital curation.

Ray asserts that the challenge is ongoing and too complex to be limited to one domain as economies of scale and international cooperation are necessary to preserve the human record in the digital age. She is optimistic that with continued support and investment in both distributed technical infrastructure and diversified human expertise, the benefits of cyberinfrastructure will extend broadly to all fields of research and discovery, to the preservation and dissemination of knowledge, and to quality of life in the digital age.

Sanderson, R., & Sompel, H. (2010). Making web annotations persistent over time. *Proceedings of the 10th annual joint conference on Digital libraries*. doi: 1003.2643

Abstract. As Digital Libraries (DL) become more aligned with the web architecture, their functional components need to be fundamentally rethought in terms of URIs and HTTP. Annotation, a core scholarly activity enabled by many DL solutions, exhibits a clearly unacceptable characteristic when existing models are applied to the web: due to the representations of web resources changing over time, an annotation made about a web resource today may no longer be relevant to their presentation that is served from that same resource tomorrow. We assume the existence of archived versions of resources, and combine the temporal features of the emerging Open Annotation data model with the capability offered by the Memento framework that allows seamless navigation from the URI of a resource to archived versions of that resource, and arrive at a solution that provides guarantees regarding the persistence of web annotations over time. More specifically, we provide theoretical solutions and proof-of-concept experimental evaluations for two problems: reconstructing an existing annotation so that the correct

archived version is displayed for all resources involved in the annotation, and retrieving all annotations that involve a given archived version of a web resource.

Credibility. Dr. Robert Sanderson is an information scientist in the Research Library at Los Alamos National Laboratory and previously a Lecturer in Computer Science at the University of Liverpool. His research focuses on the areas of scholarly communication, especially with regards to digital humanities and large scale data mining. He was won international awards for his research, including the 2010 Digital Preservation Award and both the Vannevar Bush Best Paper award at JCDL2011 and Best Poster Award at JCDL2012. Between 2009 and 2011, he was the UIUC GSLIS Honorary Research Fellow for his interdisciplinary work in digital humanities. He is an editor of several international specifications including, most recently, the W3C Open Annotation Community Draft, IETF Memento Internet Draft, and NISO Resource Synchronization. He also has close ties with the Very Large Digital Library (VLDL) community, including working with the San Diego Supercomputer Center, UC Berkeley, Stanford, Europeana and DPLA, as well as being a founding member of the UK's National Center for Text Mining (http://public.lanl.gov/rsanderson/bio/short.html).

Herbert Van de Sompel graduated in Mathematics and Computer Science at Ghent University (Belgium), and in 2000 obtained a PhD in Communication Science there. For many years, he headed Library Automation at Ghent University. After leaving Ghent in 2000, he was Visiting Professor in Computer Science at Cornell University, and Director of e-Strategy and Programmes at the British Library. Currently, he is the team leader of the Prototyping Team at the Research Library of the Los Alamos National Laboratory. The Team does research regarding various aspects of scholarly communication in the

digital age, including information infrastructure, interoperability, digital preservation and indicators for the assessment of the quality of units of scholarly communication. Van de Sompel has played a major role in creating the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), the Open Archives Initiative Object Reuse & Exchange specifications (OAI-ORE), the OpenURL Framework for Context-Sensitive Services, the SFX linking server, the bX scholarly recommender service, and info URI. Currently, he works with his team on the Open Annotation, Memento (time travel for the Web), and ResourceSync projects (http://public.lanl.gov/herbertv/bio/).

Summary. In this paper by Sanderson and Van de Sompel, from the Los Alamos National Laboratory, the authors discuss annotations as identified by John Unsworth as being a basic function common to scholarship across all disciplines, used to organize existing knowledge and facilitate the creation and sharing of new insights. Sanderson and Van de Sompel describe digital library architectures to include services that enable users to annotate managed objects. As digital libraries transition from independent silos towards web integration, they become more in line with the architecture of the web. The problem with the web architecture is as time goes by, a resource keeps the same uniform resource identifier (URI), but its representations are likely to change. For example, clicking http://www.cnn.com/, the current version of the CNN home page is returned but when clicking that same HTTP URI tomorrow, a different version of the home page will be returned. Existing annotation frameworks do not make it easy either, as after extensive research none were found with a model that allowed for different component resources to be considered at different points in time. They instead relied on

heuristics to re-attach the annotation after the inevitable modifications to the annotated resource had occurred.

The paper explores how a web-centric annotation framework can support the necessary robustness of annotations by making use of the archived representations from the time when the resource was originally annotated, by being transparent to end users and at the same time true to the web architecture's fundamental principles. Sanderson and Van de Sompel investigate two fundamental questions in this paper using the CNN example, (a) can we use the information in an annotation to retrieve an appropriate archived version of last year's CNN homepage to which the content of the annotation is relevant, and (b) given that archived version of last year's CNN homepage, can we rediscover the annotations about it? The authors conclude that it is possible to devise a time-robust, web-centric annotation framework, and more generally, that many services that include scholarly assets that are currently implemented in terms of isolated digital libraries can be reframed in a web-centric perspective.

Waters, D. (2002). Good archives make good scholars: Reflections on recent steps toward the archiving of digital information. *Council on Library and Information Resources*, 78-95. Retrieved from http://www.clir.org/pubs/reports/pub107/waters.html

Abstract. The Council on Library and Information Resources (CLIR) and, later, the Digital Library Federation (DLF) have been exploring the topic of preserving digital information for a long time. Don Waters and John Garrett wrote their landmark report, The Preservation of Digital Information, in 1996. Six years later, what is the state of preservation of digital information? This paper looks at many institutions and organizations to understand what has been accomplished.

Credibility. Donald J. Waters is the Andrew W. Mellon Foundation's Program Officer for Scholarly Communications and Information Technology. Before joining the Foundation, he served as the first director of the Digital Library Federation (1997-1999), as associate university librarian at Yale University (1993-1997), and in a variety of other positions at the Computer Center, the School of Management, and the University Library at Yale. Mr. Waters graduated with a Bachelor's degree in American Studies from the University of Maryland, College Park in 1973. In 1982, he received his PhD in Anthropology from Yale University (http://www.mellon.org/about_foundation/staff/program-area-staff/donaldwaters).

The *Council on Library and Information Resources* (CLIR) is an independent, nonprofit organization that forges strategies to enhance research, teaching, and learning environments in collaboration with libraries, cultural institutions, and communities of higher learning. CLIR promotes forward-looking collaborative solutions that transcend disciplinary, institutional, professional, and geographic boundaries in support of the public good (http://www.clir.org/about).

Summary. In this highly referenced paper, Waters talks about boundaries and territoriality, the preservation of a shared resource. Waters opens his article with Robert Frost's poem, The Mending Wall, and uses it as an analogy for the digital world. Waters introduces the topic by saying that the library, publisher, and scholarly communities are engaged in efforts to resolve the problems associated with preserving a different common resource, digital information. That information is critical for libraries and other institutions that have the responsibility of maintaining such data. Waters reports on the task force created on archiving of digital information and examines its findings. He

suggests that a serious investment is needed in archiving because of the danger of losing the cultural memory.

Waters acknowledges the work of other groups and reports, yet laments the fact that the vision is not as clear as it should be. He then proceeds to highlight the work of the Mellon Electronic Journal Archiving Program and their research. Waters spends a lot of effort addressing the political economy of public goods and the tragedy of the commons. Given the huge free-riding problem associated with the maintenance of public goods, what are the alternatives? Waters declares that it is no accident that so many of the archiving projects currently in effect are government funded. He then discusses organizational options such as government control, private interest, communities of mutual interest, publisher based archives, and research libraries in large universities. Waters discusses the economic model around archiving and how to get self-sustainability by encouraging the archiving of scholarly electronic journals. He references the Lots Of Copies Keep Stuff Safe (LOCKSS) model and JSTOR as examples of models in use, and concludes that good archives make good scholars, and if we accept the proposition that a free society depends on an educated citizenry, it is not a great leap of logic to conclude further that good archives make good citizens.

Yager, T. (2006). Technology with no past. *InfoWorld, 28*(39), 18. Retrieved from http://www.infoworld.com/d/developer-world/technology-no-past-906

Abstract. The article discusses the success of information technology, which in the future will be measured by accessibility. The term *information technology* is considered as a shorthand for electronic devices that aid humans in storage and sharing of, analysis of, protection of and access to significant amount of digitized content. Accessibility refers

to the ability of users to touch their content. Information technology access also empowers every person across the world by becoming a part of life, governance, public safety and commerce.

Credibility. Tom Yager is Chief Technologist of the InfoWorld Test Center and a senior contributing editor at InfoWorld. He is also a web developer and a Mac expert. *InfoWorld* is the leading source of information on emerging enterprise technologies, and the only brand that explains to senior technology decision makers how these technologies work, and how they can use them to drive their business. It provides in-depth technical analysis on key products, solutions, and technologies for sound buying decisions and business gain. InfoWorld features trusted industry columnists, a sharp focus on IT issues, and product test results and reviews backed by the renowned InfoWorld Test Center.

Summary. In this article published in InfoWorld, a leading technical periodical, Tom Yager discusses the role of information technology (IT) in providing access to data in the future. Yager declares that accessibility is the measure of successful IT, and he is thankful that competition is starting to drive down costs and working out in the consumers' favor with different companies coming out with competing technologies to save and access information.

Yager is optimistic about the future and the role IT will play in it from catching embezzlers to storm relief, however, it is not without its risks as the more accessible we become, the greater the chances someone will abuse technology for their own interest. Yager ends by looking back from the future and seeing IT in the same light as primates' evolution to the use of tools and weapons.

What efforts have been adopted by libraries for the purpose of archiving content to ensure (a) interoperability, (b) consistency, and the (c) safety and security of collections (Donaldson & Yakel, 2012, p. 55)?

Balas, J. (2007). By digitizing, are we trading future accessibility for current availability? *Computers in Libraries*, *27*(3), 30-32. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=24353509&site=ehost-live&scope=site

Abstract. The author comments on the concept of digitization. She provides an overview of the effort of Google to digitize the holdings and materials of some libraries. She mentions the digitization initiatives of the Open Content Alliance. She also analyzes the various issues associated with digitization, such as accessibility and availability of digital information.

Credibility. Janet L. Balas is a library information systems specialist at Monroeville, Pa. Public Library (Balas, 2007, p. 32).

Computers in libraries (CIL) is a monthly magazine that provides complete coverage of the news and issues in the rapidly evolving field of library information technology. Focusing on the practical application of technology in community, school, academic, and special libraries, CIL includes discussions of the impact of emerging computer technologies on library systems and services, and on the library community itself (http://www.infotoday.com/cilmag/).

Summary. Balas is critical of new technologies in light of the volatility of some technologies that do not last for long, however she declares that she is a technology enthusiast when it comes to technology that opens up new possibilities. Although at first

glance, it would seem that digitizing library materials to provide greater accessibility is a goal that all librarians (such as Balas) would support, yet some of the widely publicized digitization projects have been met not just with a cautious attitude from some, but also with some stiff opposition from others. Librarians whose profession is directly affected must be informed about all the issues surrounding digitization to ensure that digitization brings the most benefit to library users.

Balas examines the Google initiative to digitize the holdings of some of the world's largest libraries, and while it seems like a logical move, issues regarding copyright, legal issues, and the value of such endeavor continue to arise. One group that is not overly enthused about Google's initiative is the Open Content Alliance (OCA), which objects to Google's project due to monopoly issues. The OCA believes that if only Google holds the rights to distribute and publish a vast amount of the libraries' holdings, they wield a lot of power over the community.

Others are concerned that while having information available in digital format may make it more widely accessible now, it may actually become inaccessible in the future if the hardware and software necessary to access the information are no longer available. Digitizing materials in order to provide greater access and to preserve them seems like a great idea, as the proliferation of digitized copies makes materials available to more users and aids preservation because physical catastrophes like fires, floods, and wars can no longer destroy the only copy of any important information. However, it may not be the perfect answer just yet since future generations may not be able to access the multiple copies of digital information. In that case, all the efforts to digitize will have actually done more harm than good. Balas concludes that the challenge is to provide current

accessibility and to find a way to truly make information available to all, both in the present and in the future.

Breeding, M. (2013). Digital archiving in the age of cloud computing. *Computers in Libraries*, 22-26. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=86160357&site=ehost-live&scope=site

Abstract. The author looks at options provided by cloud computing in protecting and preserving digital assets that represent investments of time and creative energies. He notes that digital storage services provide cloud computing technologies that deliver high levels of protection. The need for proactive attention and planning for responsible care of digital data is discussed as well as the classic approach to disaster recovery planning.

Credibility. Marshall Breeding is an independent consultant, writer, and frequent library conference speaker. He is also the founder of Library Technology Guides.

Computers in libraries (CIL) is a monthly magazine that provides complete coverage of the news and issues in the rapidly evolving field of library information technology. Focusing on the practical application of technology in community, school, academic, and special libraries, CIL includes discussions of the impact of emerging computer technologies on library systems and services, and on the library community itself (http://www.infotoday.com/cilmag/).

Summary. In this paper, published in the Computers in Libraries journal, Breeding points the reader's attention to the fact of the fragility of digital information. He lists a number of risks that digital information faces, and with our increasing dependence on digital information for personal and business use, we all have a lot more to lose. In light

of that, users need to take a reasonable amount of care to protect and preserve the digital assets that represent investments in time and energy. Breeding acknowledges the efforts taken by professionals for data preservation, however he stresses that protecting the data created outside of institutional settings is the responsibility of the individual. Unfortunately, urgency in this matter is not instilled until after a disaster has taken place. Breeding covers some data security basics such as using cloud computing for backups and as a repository. He also gives some examples on different cloud solutions, but warns the user to be aware of the terms of service at all times. Breeding goes past discussion of disaster recovery and addresses digital preservation. He declares that the real challenge lies in finding ways to pass our digital assets to future generations. He briefly explains some preservation models and mentions the works of some organizations and groups that work in this space. He concludes by saying that long-term digital preservation is just as much organizational as it is technical, and the key component requires an institutional commitment to do what it takes to preserve the materials through a never-ending cycle of technology platforms. What future generations inherit from our legacies will depend on how well we tend to these principles of data security and preservation.

Careless, J. (2013). Archiving web content. *Online Searcher*, 44-46. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=86149978&site=ehost-live&scope=site

Abstract. The article features an expert view on archiving online content. Library of Congress Office of Strategic Initiative web archiving team leader Abbie Grotke told that the Library holds a project that keeps online content linked to historical events like the U.S. National Elections. Boston Public Library administration and technology director

David Leonard tells that his library keeps digital content through various storage systems like internal file servers and a Flickr storage array.

Credibility. James Careless is a freelance technology writer who has covered business processes for a number of industry-specific magazines.

Online Searcher is a journal that provides information needed to manage online research projects, assess the worth of new resources, conduct successful searches, discover pitfalls affecting information professionals, determine the utility of new technologies, and strategize services to boost their value. It features articles written by practitioner experts, as well as columns by information professionals well-known throughout the information industry (http://infotoday.stores.yahoo.net/onlinesearcher.html).

Summary. In this article, James Careless interviews three experts drawn from major library institutions at a virtual roundtable. He brings together Abbie Grotke, the web archiving team lead at the Library of Congress' Office of Strategic Initiatives, David Leonard, the Boston Public Library's director of administration and Technology, and Lauren Stokes, the virtual library manager at the Las Vegas-Clark County Library District.

Careless presents the group with a collection of questions, in an effort to find out how to preserve something reliably when it doesn't exist in physical form. Careless opens the discussion by asking the participants to speak of the different approaches each of their institutions is taking to tackle archiving of web content, streaming media, and video. He inquires about the size of each institution's collection, and the technologies they each use to store the content. The question of cloud-based storage is brought up, and Careless inquires why or why not it is being utilized at each institution. An important question

relative to this research is brought up when Careless asks his participants about how they cope with the ever-increasing change in digital storage technologies in terms of migrating to newer technologies when they are available. The responses vary from refreshing media, to data migration plans, to institutions that have not had to address this situation yet. Careless concludes with three questions regarding methods and technologies used to make archives accessible, the major challenges in archiving web content, and future plans they have for web archiving.

Gaur, R., & Tripathi, M. (2012). Digital preservation of electronic resources. *DESIDOC Journal of Library & Information Technology*, *32*(4), 293-301. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=llf&AN=78553836&site=ehost-live&scope=site

Abstract. Due to huge advances in information communication technologies (ICTs), there has been an astronomical growth of e-resources-e-journals, e-books, online databases and so on. Libraries spend phenomenally on acquisition of these e- resources as these are very popularly used by the students and researchers. Unfortunately, this growth is accompanied by many threats. Digital content (of the e- resources) is fragile and not durable. Its accessibility and use by future generations depends on technology which very rapidly evolves and changes. Hence, ensuring access of e-resources for future generation of users is a big challenge for libraries. The paper highlights various problems of digital content and elaborates how digital preservation is more demanding and challenging than preserving print copies of journals. It also gives a bird's eye view of various projects initiated for archiving digital content of scholarly journals.

Credibility. Dr. Ramesh C. Gaur is the University Librarian at Jawaharlal Nehru University (JNU), New Delhi. Prior to this, he was Librarian & Head at Kala Nidhi Division at Indira Gandhi National Centre for the arts (IGNCA), New Delhi and was the Project Coordinator of an International Bibliographic Project known as ABIA, South and Southeast Asian Art and Archeology Index. Dr. Gaur has also headed Kala Kosha Division of IGNCA during October 2009 to December 2010, and National Mission for Manuscripts (NMM), IGNCA as its Director from April 2009 to September 2009. He is the Project Coordinator for a collaborative project with C-DAC, Pune and Department of Information Technology, Government of India on "Development of Digital Repository of Indian Cultural Heritage" under the project "Centre for Excellence on Digital Preservation".

A Fulbright Scholar, Dr. Gaur holds a PhD in Library and Information Science, with exposure to several-advanced training program on the applications of IT to library management.

Summary. In this paper, Gaur discusses the increase in spending by libraries worldwide on e-resources. He defines an e-resource as any work encoded and made available by remote access and direct access. This increased spending on e-resources is due to the change in information seeking behavior among students, researchers, and faculty members. He quotes a British study that predicts that half of all serial publications will be online by 2016. The problem that is faced in this scenario is how to ensure access to all this content in the long-term. In some cases, the internet is the only medium where some of this information can be accessed, which risks the disappearance of a large number of web page and the loss of cultural and scientific data on a regular basis. As

technology changes rapidly, many mediums quickly become outdated; accordingly, preserving digital resources is a task that has to be taken seriously. This uncertainty is a major hurdle preventing libraries from moving to electronic only materials.

Gaur explains the meaning and purpose of digital preservation, then distinguishes between print journals and e-journals. He lists some issues and challenges facing digital preservation, including selection of content and use of metadata. He also lists some strategies for preservation including continuous management and maintenance, financial constraints, and collaborative efforts. Gaur also explains the role of libraries and open access repositories in preservation. He also lists some standards and efforts such as the Open Archival Information System (OAIS), the Portico digital preservation project, the LOCKSS system, CLOCKS, Pandora, KOPAL, PubMed, and others. Gaur concludes his paper by recommending that all stakeholders should come together and develop a strategy for preserving the content of electronic scholarly journals for posterity, and take measures to ensure the safety, longevity, and accessibility of collections that are least dependent on vendors and external service providers.

Meyer, L. (2009). Safeguarding collections at the dawn of the 21st Century: Describing roles & measuring contemporary preservation activities in ARL libraries. *Association of Research Libraries*, Washington, DC. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED505512

Abstract. Preservation has long been considered a fundamental responsibility of research libraries. Data on preservation activities by its members has been collected by the Association of Research Libraries (ARL) since 1987, but changing digital technologies and the research, teaching, and learning environments in which research libraries are

engaged created a need to review and examine assumptions about the types of qualitative and quantitative data needed to characterize current and emerging preservation programs. This report reflects recent shifts in libraries' content management role but is also a response to a specific recommendation of the 2006 ARL Task Force on the Future of Preservation in ARL Libraries to gather data to provide a contemporary picture of preservation programs in ARL member libraries. The report is organized around three themes: (a) preservation functions; (b) networked digital environment; and (c) collaboration. Within each section, background and analysis are provided and recommendations for consideration by ARL are posed. Overall recommendations for ways that ARL and others in the preservation community could move forward include: (a) improved cooperation and collaboration both within and among institutions to bring attention to and resolve common issues for the community; (b) ARL can encourage their members to share information about their current activities through the use of existing tools; (c) libraries should look to new ways to obtain expertise for preservation activities and use partnerships to extend their capacity; (d) all ARL libraries should maintain a core set of preservation activities appropriate for their stewardship responsibility and institutional mission; and (e) volatility of the content and technical environment requires staff to commit to continual learning. Additionally, recommendations to ARL's Statistics and Measurement Program are offered as a means for furthering conversation within the ARL and preservation community. Categories of new data to collect, the level of data needed, and specific changes for current data and definitions are suggested. Two appendixes are included: (1) List of Participants; and (2) Recommendations to the ARL Statistics & Measurement Program. A bibliography and list of acronyms are included.

Credibility. Lars Meyer is Sr. Director, Content Division at Emory University Libraries, where he is responsible for tech services, preservation, digitization, and storage.

The Association of Research Libraries (ARL) is a nonprofit membership organization comprising 125 research libraries in the US and Canada representing universities, public libraries, national libraries, and special libraries. The Association was established at a meeting in Chicago in December 1932, by the directors of 42 major university and research libraries that recognized the need for coordinated action and desired a forum to address common problems. The Association incorporated in 1961 under the laws of the District of Columbia noting that the particular business and objects of the society shall be exclusively for literary, educational and scientific purposes by strengthening research libraries.

Summary. In this paper, published in the Association of Research Libraries (ARL) journal, Meyer discusses how to safeguard digital collections in the digital age. Meyer asserts that it has long been considered a primary role for libraries to preserve content, however the changing digital technologies have created a need to review and reexamine assumptions held about the types of qualitative and quantitative data needed to characterize current and emerging preservation programs. Meyer states that preservation cannot be the sole responsibility of a single department, but it needs to be a communal effort between preservation departments and the rest of the libraries.

Meyer's report makes some recommendations regarding preserving digital content in libraries, including organizational models, developing policies, strategies, and practices, well-articulated plans for addressing fugitive digital content, not ignoring traditional preservation, staff education, establishing community practices, and others.

Meyer examines networked environments and collaboration, and concludes his paper with a final set of recommendations ranging from improved cooperation between institutions, to requiring continuous training of staff.

Moore, R. & Marciano, R. (2005). Prototype preservation environments. *Library Trends, 54*(1), 144-162. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=aph&AN=19601037&site=ehost-live&scope=site

Abstract. The Persistent Archive Testbed and National Archives and Records Administration (NARA) research prototype persistent archive are examples of preservation environments. Both projects are using data grids to implement data management infrastructure that can manage technology evolution. Data grids are software systems that provide persistent names to digital entities, manage data that are distributed across multiple types of storage systems, and provide support for preservation metadata. A persistent archive federates multiple data grids to provide the fault tolerance and disaster recovery mechanisms essential for long-term preservation. The capabilities of the prototype persistent archives will be presented, along with examples of how the capabilities are used to support the preservation of email. Web crawls, office products, image collections, and electronic records.

Credibility. Reagan W. Moore is director of the Data Intensive Cyber Environments Center (DICE Center) and a professor in the School of Library and Information Science at the University of North Carolina at Chapel Hill. He is a Chief Scientist for Data Intensive Cyber Environments at the Renaissance Computing Institute (RENCI), and President of the Data Intensive Cyberinfrastructure Foundation, which supports the open

source community for the Integrated Rule-Oriented Data System (iRODS). An internationally recognized expert, Moore coordinates research efforts in development of data grids, digital libraries, and preservation environments. Current research activities include the use of data grid technology to automate execution of management policies and validate trustworthiness of repositories. His research is funded by the National Science Foundation under the Office of Cyberinfrastructure (OCI) program and the National Archives and Records Administration, under the Electronics Records Administration Research program, and other agencies. Previous positions include Associate Director for Data Intensive Computing, Director of the Knowledge and SRB Lab, and Manager of Production Systems at UC San Diego's San Diego Supercomputer Center, and computational plasma physicist at General Atomics (https://www.irods.org/index.php/Reagan_Moore).

Richard Marciano is a Professor and Director at Sustainable Archives & Leveraging Technologies group (SALT). He holds a BS in Electrical Engineering and Avionics from the National School of Civil Aviation in Toulouse, France, and an MS and PhD in Computer Science from the University of Iowa. He specializes in Digital libraries, archives and records management, policy-based cyberinfrastructure, digital preservation, and digital humanities (http://sils.unc.edu/people/faculty).

Library Trends is a quarterly journal published by the Johns Hopkins University Press that explores critical trends in professional librarianship, including practical applications, thorough analysis and literature reviews. Both practicing librarians and educators use Library Trends as an essential tool in their professional development and continuing education. Each issue is devoted to a single aspect of professional activity or interest. In-

depth, thoughtful articles explore important facets of the issue topic. Every year, Library Trends provides breadth, covering a wide variety of themes, from special libraries to emerging technologies. An invaluable resource to practicing librarians and educators, the journal is an important tool that is utilized for professional development and continuing education (http://muse.jhu.edu/journals/lib/).

Summary. In this paper, Moore and Marciano discuss prototype preservation environments. The authors detail the efforts of the San Diego Supercomputer Center (SDSC) along with National Archives and Records Administration (NARA) to develop a prototype persistent archive. The original goal was to include an assessment of mechanisms for management of technology' obsolescence and the ability to migrate electronic records to new storage systems or *infrastructure independence*. The preservation system should be extensible and be able to use more cost-effective storage technologies as they become available. Another goal of the collaboration was to assess the scalability mechanisms that would enable the support for archives holding hundreds of millions of files and hundreds of terabytes of data. The data management technology that meets these goals is called a *data grid*.

Moore and Marciano define preservation as the process of migrating a digital entity forward in time while preserving its authenticity and integrity. A digital entity can be an electronic record, a data file created by a scientific application, a text file created by a word processing system, an image taken by a remote sensor, or any string of bits that can be named. Preservation requires the extraction of the digital entity from its environment and importing it into the preservation environment. The authors list some preservation challenges and some extraction requirements that supporting environments impose, such

as specifying the location by the archivist, persistent naming conventions, management of file properties, persistent identifiers, and persistent management of the access controls. The authors explain the concept of data grids, defined as the software infrastructure that implements a collection-based data management infrastructure for distributed data, and describe their use and function in great detail, along with preservation environment infrastructure and generic preservation environments. They also give examples of preservation environments, including The NARA research prototype permanent archive and the NHPRC Persistent Archive. In summary, the authors close by stating that the desire to support infrastructure independence, the ability to preserve digital entities as a collection, and the ability to migrate the collection to new choices for storage and database technology, are strong driving forces for the community. Data grid technology provides this capability and has been shown to scale to the size of digital holdings that are now being considered for preservation.

Rothenberg, J. & Hoorens, S. (2010). Enabling long-term access to scientific, technical and medical data collections. *RAND Europe*, Cambridge, U.K. Retrieved from http://www.rand.org/content/dam/rand/pubs/technical_reports/2010/RAND_TR567.pdf

Abstract. The British Library (BL) is considering its potential role in the intake, curation, archiving and preservation of selected scientific, technical and medical (STM) reference datasets, with the aim of providing access to and manipulation of these datasets for research purposes. In order to develop an appropriate strategy, or a set of alternative tactics, to fulfil this mission, the BL requires an analysis of the characteristics and uses of such reference data collections. The BL commissioned RAND Europe to perform a

scoping study that investigates its potential role in facilitating access to relevant datasets in the biosciences and environmental science.

This document presents the scope, approach, and findings of that study and recommendations for further action and research. The results are based on RAND's expertise and previous experience in this area and on preliminary investigation of a small set of potential candidate reference data collections.

The report is directly intended to inform the STM strategy at the British Library, but it should also be of interest to decision-makers at other national and research libraries faced with the challenges of the dynamic and complex STM landscape. It will also be relevant to other stakeholders in the research process, including researchers, funders, and organizations hosting, maintaining or governing data collections.

Credibility. Jeff Rothenberg is a senior research scientist of the RAND Corporation (Rothenbeg, 1999, p. iv).

The *Council on Library and Information Resources* (CLIR) is an independent, nonprofit organization that forges strategies to enhance research, teaching, and learning environments in collaboration with libraries, cultural institutions, and communities of higher learning. CLIR promotes forward-looking collaborative solutions that transcend disciplinary, institutional, professional, and geographic boundaries in support of the public good (http://www.clir.org/about).

Summary. In this paper, Rothenberg discusses the challenge in ensuring that digital information can be readable in the future. He claims that digital documents are vulnerable to loss via the decay and obsolescence of the media on which they are stored, and they become inaccessible and unreadable when the software needed to interpret

them, or the hardware on which that software runs, becomes obsolete and is lost. Rothenberg calls for substantial new investments and efforts to ensure the preservation of digital documents because the scope of the problem extends beyond the traditional library domain into such things as government records, environmental and scientific baseline data, documentation of toxic waste disposal, medical records, corporate data, and electronic-commerce transactions. His report explores the technical depth of the problem and analyzes the inadequacies of a number of ideas that have been proposed as solutions. Rothenberg identifies that the long-term digital preservation problem calls for a long-lived solution that does not require continual heroic efforts or repeated invention of new approaches every time formats, software or hardware paradigms, document types, or recordkeeping practices change. He proposes an extensible approach that can handle current and future documents of unknown types in a uniform way, while being flexible to evolve when needed. Rothenberg claims that most of the solutions presented in the past are not feasible due to being labor-intensive and ultimately incapable of preserving digital documents in their original forms. Rothenberg's solution is to run the original software under emulation on future computers. Implementing of this emulation approach requires: (a) developing generalizable techniques for specifying emulators that will run on unknown future computers and that capture all of those attributes required to recreate the behavior of current and future digital documents; (b) developing techniques for saving the metadata needed to find, access, and recreate digital documents, so that emulation techniques can be used for preservation; and (c) developing techniques for encapsulating documents, their attendant metadata, software, and emulator specifications in ways that ensure their cohesion and prevent their corruption.

Smith, M., & Moore, R. (2007). Digital archive policies and trusted digital repositories. In Proceedings of the *2nd International Digital Curation Conference: Digital Data Curation in Practice*. Glasgow, UK. doi:10.2218/ijdc.v2i1.16

Abstract. The MIT Libraries, the San Diego Supercomputer Center, and the University of California San Diego Libraries are conducting the PLEDGE Project to determine the set of policies that affect operational digital preservation archives and to develop standardized means of recording and enforcing them using rules engines. This has the potential to allow for automated assessment of "trustworthiness" of digital preservation archives. We are also evaluating the completeness of other efforts to define policies for digital preservation such as the RLG/NARA Trusted Digital Repository checklist and the PREMIS metadata schema. We present our results to date.

Credibility. MacKenzie Smith is Research Director at the MIT Libraries, where she oversees the Libraries' digital library research and development program. Her research focuses on the Semantic Web for scholarly communication and digital data curation, including long-term data preservation and archiving. She was the Project Director for MIT's collaboration with Hewlett-Packard to build DSpace, the open source digital archive platform now in widespread use, and has led many other research projects that advanced the international digital library agenda. Prior to MIT, MacKenzie managed the Harvard University Library's Digital Library Program Manager and held IT positions at Harvard and the University of Chicago (http://www.ka-connect.com/talks.php?vdx=93&vct=0&vcf=2).

Reagan W. Moore is director of the Data Intensive Cyber Environments Center (DICE Center) and a professor in the School of Library and Information Science at the

University of North Carolina at Chapel Hill. He is a Chief Scientist for Data Intensive Cyber Environments at the Renaissance Computing Institute (RENCI), and President of the Data Intensive Cyberinfrastructure Foundation, which supports the open source community for the Integrated Rule-Oriented Data System (iRODS). An internationally recognized expert, he coordinates research efforts in development of data grids, digital libraries, and preservation environments. Current research activities include the use of data grid technology to automate execution of management policies and validate trustworthiness of repositories. His research is funded by the National Science Foundation under the Office of Cyberinfrastructure (OCI) program and the National Archives and Records Administration, under the Electronics Records Administration Research program, and other agencies. Previous positions include Associate Director for Data Intensive Computing, Director of the Knowledge and SRB Lab, and Manager of Production Systems at UC San Diego's San Diego Supercomputer Center, and computational plasma physicist at General Atomics (https://www.irods.org/index.php/Reagan_Moore).

The *International Journal of Digital Curation* (IJDC) is a peer-reviewed electronic journal entirely devoted to papers, articles and news items on the curation of digital objects and related issues (http://www.dcc.ac.uk/resources/curation-journals/ijdc).

Summary. In this paper, Smith and Moore of the MIT Libraries, in collaboration with the University of California, San Diego, and funded by the US National Archives and Records Administration (NARA), investigate the various policies in use by operational digital archives. They identify and categorize those policies and define associated rules and state information to make them machine encodable and, wherever possible,

enforceable. The authors map the PoLicy Enforcement in Data Grid Environments (PLEDGE) Project policies for enterprise, archive, collection, and item levels, evaluate what is missing, and attempt to demonstrate the set of rules that automatically validate the trustworthiness of a repository.

Smith and Moore also observe the mapping of assessment criteria to the management policies and postulate how a preservation environment might be assessed to insure that the system is complete. They conclude with the assertion that it is possible to develop preservation systems that are subject to rigorous assessment, and believe that this will allow preservation environment to scale appropriately in the coming decades.

Watry, P., Larson, R., Sanderson. (2006). Knowledge generation from digital libraries and persistent archives. *Research and Advanced Technology for Digital Libraries, 4172,* 504-507. Retrieved from http://cgi.csc.liv.ac.uk/~azaroth/papers/ecdl2006.pdf

Abstract. This poster describes the ongoing research of the Cheshire project with a particular focus on knowledge generation and digital preservation. The infrastructure described makes use of tools from computational linguistics, distributed parallel processing and storage, information retrieval and digital preservation environments to produce new knowledge from very large scale datasets present in the data grid.

Credibility. Paul Watry is the Deputy Director of the National UK Text Mining Centre (NaCTeM) and Principal Investigator for the Multivalent digital preservation architecture project and the Cheshire digital library system. His primary area of interest is in computational linguistics and in bibliographic analysis. A core activity is to develop and implement a strategy, which will embrace both electronic and traditional information resources and address the needs of both research and learning (http://www.liv.ac.uk/psychology-health-and-society/staff/paul-watry/).

Dr. Larson specializes in information retrieval and database systems, with an emphasis on the system internals. He was involved in the design and development of UC public access online union catalog (MELVYL). He also helped design the algorithms used in the Inktomi web search engine. He is the principal designer of the Cheshire information retrieval system, and active in international IR evaluations including cross-language evaluations like CLEF and NTCIR. His focus is on Information retrieval system design and evaluation (http://www.ischool.berkeley.edu/people/faculty/raylarson).

The *10th European Council on Research and Advanced Technology for Digital Libraries* (ECDL) was held in Alicante (Spain) in September 2006. The ECDL has become the major European conference on digital libraries, and associated technical, practical, and social issues, gathering researchers, developers, content providers and users in the field (http://www.ecdl2006.org/index.html).

Summary. In this paper, written jointly by the University of Liverpool and the University of California, Berkeley, the authors discuss working on technologies and infrastructures that will support digital library services and persistent archives based on the Storage Resource Broker (SRB) data grid technology. The authors cover the background and rationale behind the concept, including challenges facing the scientific community and the requirements for dealing with this type of data.

The authors also list some components of persistent archives, including computational linguistics, digital library technologies, presentation technologies, and data grid technologies. Their strategy is to generate knowledge across multiple domains by combining data grid abstractions for storage (using the SRB) with presentation applications (Multivalent) and digital library and content management functionalities

(Cheshire). The authors realize the need to guarantee the manipulation of digital entities in the future across the different infrastructures, which may be used in different scientific disciplines. Their work helps facilitate digital preservation, thus inspiring a new approach to knowledge generation which will in the future make intelligent use of the massive data stores characteristic of the scientific world.

Conclusion

This annotated bibliography compiles and summarizes the works of thirty-one references from different authors and selected from peer-reviewed publications, professional journals, and conference proceedings. The references present ongoing discussions about digital archiving and data preservation in three sub-topic areas: (a) archiving practices in private and public organizations involved in commerce, science, and engineering for disaster recovery and long-term accessibility; (b) best practices to manage data preservation; and (c) efforts adopted by libraries for the purpose of archiving content to ensure interoperability, consistency, and the safety and security of collections (Donaldson & Yakel, 2012, p. 55).

The discussions exist within a contextual definition of digital archiving. Ludascher (2001) provides a fairly detailed definition of digital archiving as: capturing and preserving information so it can be discovered, rediscovered, accessed, and presented at any time in the future (p. 54). The definition provided by Goth (2012) is much simpler, and simply states that digital preservation means keeping artifacts useful across space and time (p. 11).

Discussions about Digital Archiving and Data Preservation Practices

The selected authors presented in this sub-topic agree that the continued preservation and access of digital information cannot be guaranteed (Chen, 2001, p. 2). Accordingly there is a need to develop policies and practices for information governance to define data access and information retention (Gantz, 2008, p. 2).

A range of factors and problems are presented for consideration regarding digital archiving and information preservation practice. Chen (2001) asserts that *cheaper storage* that allows creation and storage of more digital content puts increasing pressure on those responsible for developing strategies for storing, retaining, and purging information on a regular basis (p. 4).

Rabinovici et al. (2011) predict a growing number of organizations will *require the preservation of large volumes of digital content*, including the need to maintain access to it for reasons such as sustainability of business assets, retention of intellectual property, and appreciation of cultural and scientific history, in addition to regulatory compliance (p. 1). Moore, Rajasekar, and Wan (2005) state that many organizations and entities are faced with the problem of *organizing digital entities into collections* and *assigning descriptive metadata to support discovery* (p. 578).

Ludascher, Marciano, and Moore (2001) present the challenge of archiving digital content given *the limited lifetime of storage* and the technological obsolescence of infrastructure (p. 54). Moore (2008) goes further by associating the concept of preservation with communication with the future. He defines the major challenges of incorporating new technology effectively, while conserving preservation properties such as authenticity, integrity, chain of custody, the ability to characterize how prior preservation processes have been controlled by preservation management policies, and the ability to verify that policies and standards are working properly (p. 64).

These selected authors propose various solutions to the problems they present. Ludascher (2001) presents *a framework for the preservation of digital data* based on a forward migration approach using XML (p. 62). Chen (2001) advises decision makers to *consider the costs of making information available to communities worldwide via the internet*, and describes a compelling need to meet the research challenge to resolve the conflict between the creation context and the use of context to facilitate digital information preservation (p. 6).

Rabinovici et al. describe *the standardization work of SIRF*, a long-term storage container format that provides strong encapsulation of large quantities of metadata together with the data at the storage level, and allows easy migration of the preserved data across storage devices. SIRF

also enables a mountable storage container to be self-describing and self-contained to the extent possible (p. 9).

Moore (2008) introduces *a theory of preservation* that manages communication from the past, while communicating with the future, and demonstrates *rule-based data grids* that can verify that prior policies correctly enforced preservation properties (p. 63). Moore, Rajasekar, and Wan (2005) introduce data grids with an ability to manage the consistency of federated data collections while flowing information and data from digital libraries through grid services into preservation environments (p. 586). Lastly, Gantz (2008) proposes three imperatives organizations face while dealing with the *explosion of the digital universe* and the tools in place to tame those challenges, including: (a) the need to transform their existing relationships with the business units, (b) the need to spearhead the development of organization wide policies for information governance, and (c) the need to rush new tools and standards into the organization (p. 2).

Discussions about Standards and Best Practices to Manage Data Preservation

Sub-topic 2 investigates the problems and potential solutions involved in defining best practices and standards, both needed by and currently available to the preservation community. The selected authors each present a different aspect of the larger discussion.

The National Science Board (NSB) (2005) points to the rapid multiplication of collections with a potential for decades of curation (p. 9). McClure (2006) acknowledges the *fallacy* that if something is on the web, it will stay there, and the daunting task of preserving the seemingly ephemeral web content (p. 14). Maniatis et al, (2005) state that there is no guarantee of long term access to digital content (p. 3). On the other hand, Liu (2003) states that the creation of an institutional guarantee for trusted digital preservation is instrumental to increasing

people's confidence and trust in digital media, since there are no precursors for preserving documents of this nature (p. 94).

McGath (2013) reviews several noteworthy formats that have attempted digital preservation, but with limited success (p. 1). Gladney (2004) claims that the immense investments in creating and disseminating digitally represented information have not been accompanied by commensurate effort to ensure the longevity of information of permanent interest (p. 1), and Goth (2012) claims that while data is expanding at an unprecedented rate, funding for research in digital preservation must be increased and policy coordination needs improvement (p. 11).

The selected authors also propose solutions for standards and best practices in digital preservation. Ray (2012) proposes an *interconnected network of trustworthy repositories, well-curated data, and a diverse body of experts* who contribute to the preservation and discoverability of digital content with long-term value (p. 620). Berman (2008) proposes ten guidelines for data stewardship including *planning, awareness, transition plans, and making multiple copies of data* (p. 56). Sanderson and Van de Sompel (2010) recommend combining the temporal features built into the emerging Open Annotation model, with the capability offered by the recently introduced Memento framework which allow HTTP-navigation from the URI of a resource to archived versions thereof (p. 9). Donaldson and Yakel (2012) advocate the value of using theoretical frameworks from DOI and MIS to better understand the social and technical issues archivists face while implementing technology standards (p. 80).

McGath advises that ongoing efforts by many different people working independently will be necessary to keep up with the growing variety of formats and standardization (p. 7). Goth (2012) asserts that any data-management application consists of policies applied to validate

the preservation process, despite the lack of national-level coordination of data preservation standards (p. 13). Gladney (2004) articulates principles for a Trustworthy Digital Object (TDO) design that addresses every technical problem and requirement articulated in the literature. Its central elements are an encapsulation scheme for digital preservation objects and encoding using extended Turing-complete virtual machines (p. 6).

Discussions within Library Science about Archiving Content

Libraries have historically put much effort into archiving and data accessibility, thus this work deserves specific review. In the International Journal of Digital Curation, Smith and Moore (2007) investigate the various policies in use by operational digital archives, and identify and categorize those policies (p. 93). In the same context, Rothenberg's (2010) study aims to assist the British Library (BL) in developing an appropriate strategy for the intake, curation, archiving, and preservation of scientific, technical, and medical (STM) reference datasets, in order to provide access to these datasets for research purposes (p. xv).

Problems that face libraries are explored in depth by multiple authors; Gaur and Tripathi (2012) discuss the problems of digital content in light of the increase in spending by libraries on e-resources (p. 293). Meyer (2009) identifies the problem of changing digital technologies and the research, teaching, and learning environments in which research libraries are engaged, which create a need to review and examine assumptions about the types of qualitative and quantitative data needed to characterize current and emerging preservation programs (p. 7). Moore and Marciano (2005) introduce some challenges facing preservation when attempting to extract a digital entity from its supporting environment (p. 145).

Authors provide some solutions to problems facing libraries. Breeding (2013) investigates protecting data at the personal, professional, or institutional level through the use of

digital storage services that provide thorough cloud computing technologies to deliver extraordinarily high levels of protection (p. 22). Watry, Larson, and Sanderson (2006), through their work with the University of Liverpool and the San Diego Supercomputer Center (SDSC), introduce tools to address the challenges of generating knowledge, which are compounded and facilitated by the large amount of data currently being generated by the scientific community (p. 2). Careless (2013), through his work interviewing library professionals, uncovers how to preserve something reliably that doesn't exist in a physical form, specifically in light of the web playing an ever-increasing role in human lives (p. 44).

The authors also provide various methods to disentangle the problems they list in their research. Meyer (2009) identifies a range of new preservation activities and highlights key challenges faced by preservation programs in the shorter term. He informs library and preservation leaders of developments in preservation self-assessment tools for member libraries (p. 41). Moore and Marciano (2005) propose data grid technologies to provide preservation environment capabilities, which have been shown to scale to the size of digital holdings that are now being considered for preservation (p. 160). Gaur and Tripathi (2012) advocate that libraries take measures for ensuring the safety, longevity, and accessibility of collections and that they should be least dependent on vendors and external service providers for those tasks (p. 299).

Carless (2013) summarizes a librarian round-table by recommending that they continue to work with their partners to tackle the shared challenges of preservation and focus on launching digital repositories while sticking with the basics of moving forward with digitizing content (p. 46). Similarly, Smith and Moore (2007) present an end-to-end description of the management properties needed in a preservation environment, from assessment criteria to the rules that express the management policies and the descriptive and technical metadata needed to validate

the assessment results (p. 100). Lastly, Watry, Larson, and Sanderson's (2006) strategy is to generate knowledge across multiple domains, by combining data grid abstractions for storage with presentation applications and digital library and content management functionalities to guarantee the manipulation of digital entities in the future across the different infrastructures which may be used in different scientific disciplines (p. 4).

References

Balas, J. (2007). By digitizing, are we trading future accessibility for current availability? *Computers in Libraries, 27*(3), 30-32. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=24353509&site=ehost-live&scope=site

Bell, C., & Frantz P. (2012). Critical Evaluation of information sources. *University of Oregon Libraries.* Retrieved from https://library.uoregon.edu/guides/findarticles/credibility.html

Berman, F. (2008). Got data?: A guide to data preservation in the information age. *Communications of the ACM* 51, 50-56. doi:10.1145/1409360.1409376

Breeding, M. (2013). Digital archiving in the age of cloud computing. *Computers in Libraries,* 22-26. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=86160357&site=ehost-live&scope=site

Busch, C., De Maret, P. S., Flynn, T., Kellum, R., Le, S., Meyers, B., Saunders, M., Palmquist. M. (2013). Content Analysis. Writing@CSU. *Colorado State University*. Retrieved from http://writing.colostate.edu/guides/guide.cfm?guideid=61

Campos, R. (2007). Digital libraries and engines of search: New information systems in the context of the digital preservation. 2007. *Euro American conference on Telematics and information systems,* 9. doi: 10.1145/1352694.1352703

Careless, J. (2013). Archiving web content. *Online Searcher*, 44-46. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=cph&AN=86149978&site=ehost-live&scope=site

Center for Research Libraries (2007). Trustworthy repositories audit and certification:

Criteria and checklist. *Online computer library center*. Retrieved from www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Chen, S.-S. (2001). The paradox of digital preservation. *Computer,* 24-28. doi: 10.1109/2.910890

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: SAGE Publications.

Donaldson D., & Yakel, E. (2012). Secondary adoption of technology standards: The case of PREMIS. *Archival Science, 13*(1). doi: 10.1007/s10502-012-9179-0

Gantz, J. (2008). The diverse and exploding digital universe. *IDC Whitepaper*. Retrieved from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Gaur, R., & Tripathi, M. (2012). Digital preservation of electronic resources. *DESIDOC Journal of Library & Information Technology, 32*(4), 293-301. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=llf&AN=78553836&site=ehost-live&scope=site

Gladney, H. (2006). Principles for digital preservation. *Communications of the ACM, 49*(2), 111-116. doi: 10.1145/1113034.1113038

Goth, G. (2012). Preserving digital data. *Communications of the ACM, 55*(4), 11-13. doi:10.1145/2133806.2133811

Guthrie, K. (2002). Challenges and opportunities presented by archiving in the electronic era. The Johns Hopkins University Press, Retrieved from http://muse.jhu.edu.libproxy.uoregon.edu/journals/pla/

Hart, P. & Liu, Z. (2003). Trust in the preservation of digital information. *Communications of the ACM 46*, 6 93-97. doi:10.1145/777313.777319

Heutelbeck, D., & Klas, C. (2012). Ijdl focused issue on persistent archives. *International Journal on Digital Libraries, 12*(1), 1. doi: 10.1007/s00799-012-0093-0

Hewitt, M. (2002). Carrying out a literature review. *Trent focus group*. Retrieved from https://ce.uoregon.edu/aim/Capstone07/HewittLitReview.pdf

Jackson, S., Edwards, P., Bowker, G., & Knobel, C. (2007), Understanding infrastructure: history, heuristics, and cyberinfrastructure policy. *First Monday, 12*(6) Retrieved from http://firstmonday.org/issues/issue12_6/index.html

Kenney, A., & Stam, D. (2002). The state of preservation programs in American college and research libraries: Building a common understanding and action agenda. Optimizing collections and services for scholarly use. *Association of Research Libraries,* Washington, DC. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED481030

Kuny, T. (1997). A digital dark ages? Challenges in the preservation of electronic information. *63RD IFLA Council and General Conference*. Retrieved from http://archive.ifla.org/IV/ifla63/63kuny1.pdf

Lorie, R. A. (2001). Long term preservation of digital information. *1st ACM/IEEE-CS joint Conference on Digital libraries,* 346-352. doi:10.1145/379437.379726

Lose the paper clutter, not the client. (2006). *Practical Accountant, 39*.20. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=bth&AN=22232194&site=ehost-live&scope=site

Ludascher, B., Marciano, R., & Moore, R.. (2001). Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *Sigmod Record, 30*(3), 54-63. doi:10.1145/603867.603876

Maniatis, P., Roussopoulos, M., Giuli, T., Rosenthal, D. & Baker, M. (2005). The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on Computer Systems, 23*(1), 2-50. doi: 10.1145/1047915.1047917

McGath, G. (2013). The format registry problem. *Code4Lib Journal*, 1-10. Retrieved from http://journal.code4lib.org/articles/8029

Meyer, L. (2009). Safeguarding collections at the dawn of the 21st Century: Describing roles & measuring contemporary preservation activities in ARL libraries. *Association of Research Libraries*, Washington, DC. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED505512

Moore, R. (2008). Towards a theory of digital preservation. *The International Journal of Digital Curation*, 63-75. doi:10.2218/ijdc.v3i1.42

Moore, R. & Marciano, R. (2005). Prototype preservation environments. *Library Trends, 54*(1), 144-162. Retrieved from http://search.ebscohost.com.libproxy.uoregon.edu/login.aspx?direct=true&db=aph&AN=19601037&site=ehost-live&scope=site

Moore, R., Rajasekar, A., & Wan, M. (2005). Data grids, digital libraries, and persistent archives: An integrated approach to sharing, publishing, and archiving data. *Proceedings of the IEEE, 93*(3), 578-588. doi:10.1109/JPROC.2004.842761

National Science Board (2005). Long-lived digital data collections: Enabling research and education in the 21st century. *National science foundation*. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

Rabinovici, S., Baker, M., Cummings, R., Fineberg, S., & Marberg, J. (2011). Towards SIRF:

Self-contained information retention format. *Proceedings of the 4th Annual International Conference on Systems and Storage*. New York: ACM. doi: 10.1145/1987816.1987836

Ray, J. (2012). The rise of digital curation and cyber infrastructure: From experimentation to implementation and maybe integration. *Library Hi Tech, 30*(4), 604-622. doi: 10.1108/07378831211285086

Rothenberg, J. & Hoorens, S. (2010). Enabling long-term access to scientific, technical and medical data collections. *RAND Europe*, Cambridge, U.K. Retrieved from http://www.rand.org/content/dam/rand/pubs/technical_reports/2010/RAND_TR567.pdf

Sanderson, R., & Sompel, H. (2010). Making web annotations persistent over time. *Proceedings of the 10th annual joint conference on Digital libraries*. doi: 1003.2643

Smith, M., & Moore, R. (2007). Digital archive policies and trusted digital repositories. In Proceedings of the *2nd International Digital Curation Conference: Digital Data Curation in Practice*. Glasgow, UK. doi:10.2218/ijdc.v2i1.16

Standards at the library of congress. (2012). Retrieved from http://owl.english.purdue.edu/owl/resource/560/10.

University of North Carolina. (2012.). Literature reviews. *University of North Carolina at Chapel Hill*. Retrieved from http://www.unc.edu/depts/wcweb/handouts/literature_review.html

Waters, D. (2002). Good archives make good scholars: Reflections on recent steps toward the archiving of digital information. *Council on Library and Information Resources*, 78-95. Retrieved from http://www.clir.org/pubs/reports/pub107/waters.html

Watry, P., Larson, R., Sanderson. (2006). Knowledge generation from digital libraries and

persistent archives. *Research and Advanced Technology for Digital Libraries, 4172*, 504-507. Retrieved from http://cgi.csc.liv.ac.uk/~azaroth/papers/ecdl2006.pdf

Yager, T. (2006). Technology with no past. *InfoWorld, 28*(39), 18. Retrieved from http://www.infoworld.com/d/developer-world/technology-no-past-906