Finding Balance Between Archival Principles and Real-Life Practices in an Institutional Repository

ERIN O'MEARA and MEG TUOMALA

RÉSUMÉ Les archivistes d'aujourd'hui ont beaucoup de mal à trouver un juste équilibre entre la théorie et la pratique dans leurs tâches professionnelles, surtout quand ils doivent conceptualiser et mettre en place un dépôt d'archives institutionnel. Cet article explore les croisements entre la pratique par le biais d'une analyse de la théorie archivistique pertinente et des conseils réalistes, puis il examine comment certaines de ces théories ont servi ou pas au développement de la *Carolina Digital Repository (CDR),* le dépôt d'archives institutionnel de la University of North Carolina à Chapel Hill (UNC).

ABSTRACT Today's archivists struggle to find a balance between theory and practice in their professional duties, especially when tasked with designing and implementing an institutional repository. This article explores the intersections between theory and real-life practice through a discussion of relevant archival theory and realistic advice, and an examination of how some of these theories were, or were not, applied in the development of the Carolina Digital Repository (CDR), the institutional repository at the University of North Carolina at Chapel Hill (UNC).

Introduction

Is it possible to design and implement a preservation repository founded on traditional archival theory and principles? This is a question that electronic records archivists struggle with as they attempt to find a balance between archival principles, and practices in their professional duties. Although we can speak only from our own experiences, this article attempts to answer this question through a discussion of relevant archival principles and practical advice, and an examination of how some of these principles were, or were not, applied in the development of the Carolina Digital Repository (CDR), the institutional repository at the University of North Carolina at Chapel Hill (UNC). We close with a discussion of how the institutional repository[1] development process at UNC has shaped plans for the future, ideas and hopes for improvements, and how the experience can inform larger issues surrounding the development of preservation repositories.

Archival Principles and Practices

In the archival profession most practical decisions are based on theory and principles. This section explores both traditional and new appraisal and preservation principles and theories that can readily be applied to the preservation of electronic records and digital archival materials within the context of the development of an institutional repository. It also addresses some of the recent literature that specifically focuses on issues archivists should consider when developing an institutional repository.

Traditional Approaches

Sir Hilary Jenkinson's *A Manual of Archive Administration*, originally published in 1922, outlines the history and functions of archives and provides explanations for, and best practices on, an array of archival topics such as the role and duties of archivists, the appropriate selection and acquisition of records, and the idea and importance of custody in recordkeeping. Early in this seminal work, Jenkinson lays out the fundamental duties of an archivist: firstly to "safeguard" the records in his or her custody (and their essential qualities), while secondarily providing access to these records.[2] Jenkinson is explicit that the order of these roles is of utmost importance and should not be reversed. This principle is strongly reflected in some archival schools of thought, and is the approach adopted for the current incarnation of the CDR.

Jenkinson also discusses custody, and its importance within archives and recordkeeping. He describes cases where there is a need to transfer records to another steward, such as when there are records for an organization that no longer exists. This transfer can be done "without the archives losing their character ... [as long as] the chain of custody remains unbroken."[3] This principle is especially relevant within an electronic records context where the custody and state of records can sometimes be in flux. Unlike analogue documents, electronic records are not fixed to a medium; digital materials can move between computers and storage devices easily, obscuring their history and relevant context. Although he was writing long before digital technolo-

[1] The authors consider an institutional repository to be not just an open access repository for scholarly output, but also one where digital content managed and/or created by an organization can and should be deposited.

[2] Sir Hilary Jenkinson, *A Manual of Archive Administration* (London, 1922), 15.

[3] Ibid., 33–39.

gies existed, Jenkinson describes requirements for ensuring that the chain of custody is not broken when transferring records between agencies, including the transfer between creator and repository. Requirements include that the new agency follow customary rules for the management of records and ensure the continued existence of the archives; and that the new steward take the materials *in toto* without just selecting what Jenkinson calls the "pretty specimens."[4]

As a reaction to Jenkinson's model of appraisal, T.R. Schellenberg's "The Appraisal of Modern Public Records," originally published in 1956, offers many recommendations and analytical "tests" that remain useful today for archivists selecting and working with electronic records in an institutional repository. Regardless of today's very different circumstances and environments, we can learn from Schellenberg's advice that "very voluminous" records may need to be reduced before being accessioned into our repositories, and that we must take "great care" to appraise these records in order to "retain those that have value."[5] Schellenberg's tests of evidential and informational values are still applicable as we aim to preserve both digital evidence of an organization itself and its functions, as well as information on its related "people ... bodies, things, [and] problems."[6]

Schellenberg's "test of uniqueness" states that the information "need not be completely dissimilar from all other information" but that records should contain evidence for, or information about, "matters on which other documentary information does not exist as fully or as conveniently."[7] The test of uniqueness is pertinent to the context of an institutional repository, given the likelihood of duplication both within its holdings and within and across collections.

Schellenberg's "test of form," where both the form of the information held within the record and the form of the record itself is addressed, is also relevant to an institutional repository.[8] He advises archivists to seek out records "that represent concentrations of information," that are in a physical form that enables ease of use and whose arrangement "most facilitates the extracting of information."[9] The test of form is relevant to an institutional repository where material can easily be made available to researchers in its most useful form. Institutional repositories allow users to view access copies of files, such as PDFs or JPEGs, instead of extremely large and unwieldy files, such as TIFs or uncompressed audio files stored as preservation master copies.

[4] Ibid., 41.

[5] T.R. Schellenberg, "The Appraisal of Modern Public Records," *Bulletins of the National Archives* 8 (October 1956), http://www.archives.gov/research/alic/reference/archives-resources/appraisal-intro.html (accessed 28 November 2011).

[6] Ibid., 58.

[7] Ibid., 63.

[8] Ibid., 65.

[9] Ibid., 65–66.

Perhaps most significant is what Schellenberg calls the "test of importance," which calls on archivists to appraise records based on their projected significance, utility, and research value.[10] This is the most difficult and imperceptible part of appraisal because it requires the analytical skills of the archivist, and sometimes input from outside agencies and subject specialists.[11]

Although archivists do aim to preserve objectively and expansively the histories of the organizations, constituents, and cultures whose records they preserve, it is prudent to reflect on Shellenbergian appraisal standards when faced with the flood of records that today's electronic environment has encouraged. Accessions should occur only after achieving a complete assessment and understanding of the records themselves and the repository within which they will be preserved. In the case of an institutional repository, these archival appraisal guidelines can inform the collecting efforts surrounding data sets and other non-traditional, scholarly output. They can also help form a sound collection development policy, necessary in light of the large volume of electronic records archivists are facing.

While appraisal theory can help institutional repositories decide *what* to collect, the core archival concepts of authenticity and reliability can help determine *how* institutional repositories will preserve their content for the long-term. In her 1995 article, "Reliability and Authenticity: The Concepts and Their Implications," Luciana Duranti defines the core archival science concepts of reliability and authenticity, and their necessity to understanding the true nature of a record.[12] According to Duranti, *reliability* is "the authority and trustworthiness of the records as evidence, the ability to stand for the facts they are about."[13] Reliability is achieved through the *procedure of creation*, which is a way to describe and define the controls in place for creating and handling records within a recordkeeping system. Naturally, the more rigorous these procedures are and the more routinely they are used in practice, the more the reliability of the records created in the system is enhanced.[14] The *authenticity* of a record is derived from the guarantee that the record is what it purports to be, and has not undergone any alteration or falsification since its creation.[15]

Duranti's article concludes with a frank assessment of reliability and authenticity in the modern record-making and recordkeeping environment, asserting that "the easiness of electronic records creation and the level of

[10] Ibid., 66.

[11] Ibid., 67.

[12] Luciana Duranti, "Reliability and Authenticity: The Concepts and Their Implications," *Archivaria* 39 (Spring 1995): 5.

[13] Ibid., 6.

[14] Ibid.

[15] Ibid., 7–8.

autonomy that it has provided to records creators, coupled with an exhilarating sense of freedom from the chains of bureaucratic structures, procedures, and forms, have produced the sloppiest records creation ever in the history of record-making" and that "electronic records, as presently generated, might be authentic, but they are certainly not reliable."[16] It follows that in order for records to be preserved as both authentic and reliable, they must be created within a recordkeeping system that ensures that both of these elements are recognized and controlled. This level of control likely reaches far beyond the scope of most preservation-centred institutional repositories that manage content long after its active life; nevertheless the "reliable" record creator should be apprised of good records management practices and policies.[17] To increase reliability, staff can perform pre-custodial intervention and provide guidance to records creators on records management principles applied to the records identified for transfer to the institutional repository.

In their 1996 report, "The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project," Luciana Duranti and Heather MacNeil report on the findings of the University of British Columbia's Master of Archival Studies (UBC-MAS) research project, and provide a more in-depth look at how the concepts of reliability and authenticity can be maintained in the digital recordkeeping and preservation environment of the day. Additionally, they introduce the concept of *integrity* for describing both the reliability and authenticity of a record.

The UBC-MAS research project was a deductive research inquiry into how to identify methods for preserving the integrity of records created in digital form.[18] Within the context of this project, Duranti and MacNeil provide a very strict definition of a record. This definition does not exactly align with what many archivists recognize to be records being preserved as such in recordkeeping and institutional repository environments.[19] The UBC-MAS research project suggested a highly controlled recordkeeping environment

[16] Ibid., 9.

[17] Records management at UNC is focused mainly on the administrative records of the university, not on faculty papers or records documenting their research related activities. University Archives and Records Management Services, UNC's records management unit, also provides advice and assistance to student groups who wish to preserve their organizational records (with the exception of academic output material, e.g., journals and digital scholarship). Some of these activities are gaining popularity in units such as academic research computing and faculty assistance centres.

[18] Luciana Duranti and Heather MacNeil, "The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project," *Archivaria* 42 (Fall 1996): 46.

[19] The definition of a public record in many public records and freedom of information legislations includes much more than the diplomatics-based definition of a record. Examples include some publications, system log files, and other transitory documents. Certain public records that would not normally fit the diplomatics definition are also deemed permanent by the public and organizational records retention schedules.

that would employ technical actions such as the registration and classification of records.[20] These environments and actions are not always feasible in some environments such as active recordkeeping in offices, or even preservation environments within institutional repositories. According to Duranti and MacNeil, the reason for these controls – and the key to preservation – is to maintain the reliability of the record as well as the archival bond between records.[21] *Archival bond* is the context and relationships that link all of the records in a collection together, especially those immediately preceding, and subsequent to, an individual record.[22]

While Duranti and MacNeil attempt to translate these system requirements and actions into the digital environment, many institutional repositories will not be able to apply all such requirements because of varied institutional needs and finite resources. Their definition of a record is too narrow to be successfully used by many practicing archivists; however, the concept of archival bond, and the need to maintain context and relationships of records in a collection in order to preserve them fully and faithfully, is an attainable and worthy goal for institutional repositories.[23] More work needs to be done in demonstrating the contextual linkages found in repository environments. The documentation from subsequent InterPARES work can help formulate system requirements for repositories and provide guidance for records creators.

A New Paradigm

In the mid-1990s a new and sometimes controversial paradigm for understanding and dealing with electronic records emerged. Proponents of the new paradigm argued that “the very nature of electronic records requires archivists to adopt new ideas that would change or overturn traditional archival principles” such as those presented by T.R. Schellenberg and others.[24] In her article “Schellenberg in Cyberspace,” Linda J. Henry questions and refutes this new paradigm, advocating for faith in traditional archival theory and principles. In

[20] Duranti and MacNeil, 48.

[21] Ibid., 53.

[22] Ibid., 49.

[23] The InterPARES Projects followed the UBC-MAS Project and tried to apply more explicit requirements to electronic records preservation on an international level. Luciana Duranti and Randy Preston edited a large electronic publication that represented the activities and findings of the InterPARES 2 Project (*International Research on Permanent Authentic Records in Electronic Systems [InterPARES] 2: Experiential, Interactive and Dynamic Records*, [Padova: CLEUP, 2008]). The InterPARES 2 Project (2002–2007) was an extension of the first phase of the InterPARES 1 Project (1999–2001). It expanded the scope of research to include the investigation of experiential, interactive, and dynamic electronic records. The book is extremely detailed and is broken down by the various teams, domains, focus groups, and task forces that formed the project team.

[24] Linda Henry, “Schellenberg in Cyberspace,” *American Archivist* 61 (Fall 1998): 309.

particular, Henry provides specific arguments against three of the key ideas associated with the new paradigm: the effort to redefine what a record is, the concept of the records continuum, and the postcustodial model for the preservation of electronic records.

New paradigm supporters require that records, by definition, must provide evidence of business transactions, and thereby exclude personal papers and other documentary materials altogether.[25] Henry argues that this "new definition of a record is too narrow," and focuses on individual transactions that do not provide evidence of the big picture, thus ignoring many materials that may have permanent value such as "databases and personal papers."[26] She sees the new definition of a record as "an obstacle to archival work" and urges, "instead of asking whether documentary materials are records, archivists should ask if those materials are important."[27] This concept is significant in digital acquisitions, as archivists are seeing new types of digital objects (e.g., large databases, new forms of digital scholarship, and emerging formats) that challenge the notion of "recordness" even further.

A second key concept of the new paradigm is the records continuum, the idea that "there should be no distinction between archival and records management work." A records continuum replaces the life cycle concept of records that traditionally defined and delineated the responsibilities of records creators, records managers, and archivists.[28] The records continuum calls for the archivist to "hold responsibility beginning before creation, through maintenance, preservation, and use." Henry interprets this to mean that archivists would essentially "usurp the role of creator" making "records 'less genuine, less authentic', and thus sacrifice their highest virtue: neutrality."[29]

Henry also refutes the idea of postcustodialism, which supports the decentralization of archives, and envisions an environment where the creators of records take care of their own archival records. She states that an environment of non-custody would result in "records lost and damaged ... in vast disarray" and archivists left to "deal with the aftermath."[30] Henry goes on to describe the potentially "deleterious effect" of postcustodialism on archives, where historical records in active systems would be easily destroyed or changed without the creator's knowledge.[31] This could easily happen to records stored in enterprise software applications, such as student records at a university. Universities have allocated significant resources for these systems; this is an area where

[25] Ibid., 315.
[26] Ibid.
[27] Ibid., 316.
[28] Ibid., 318.
[29] Ibid., 319.
[30] Ibid., 320.
[31] Ibid.

archivists can step in and act as stakeholders and consultants, and help define system specifications and policies for use in order to ensure reliability.

While pointing out some of the problems with the new paradigm, Henry also highlights her premise that perhaps it is not impossible to deal with the problems raised by electronic records with traditional archival theory and principles after all. She states that while electronic records present archivists with new challenges, solutions will come from an examination of what we already know, not a "dismantling of archival theory and practice."[32]

While most of the writing on electronic records focuses on governmental and organizational records, Adrian Cunningham's 1999 article "Waiting for the Ghost Train: Strategies for Managing Electronic Personal Records Before It Is Too Late" addresses the issues and challenges electronic records have introduced to the field of personal records and manuscripts, and gives some suggestions for their management. Several of these issues and challenges overlap with those presented with electronic records in all fields, especially Cunningham's discussion of the tone of the dialogue surrounding electronic records, and some of the mentalities that archivists have adopted toward managing them.

While Cunningham offers some very useful suggestions on the management of personal records, especially relevant to our context is his call for a re-examination of the records continuum and the benefits of both "continuum thinking" and outreach, and some "pre-custodial intervention."[33] Cunningham asserts that the traditional division of "current records" from "historical records" is an artificial one, and that "a record is a record is a record."[34] He also states that "archivists cannot afford to be the passive recipients of records that are no longer required by their creators," and that "the traditional post hoc approach to record keeping ... is patently inadequate in the electronic environment."[35]

Cunningham suggests that while archivists can never "know what will happen in the future ... there are things about the present that we do know will be of enduring interest to society in the future," and that "we should not be derelict in our duty to the future by neglecting those people in the present who we know are significant."[36] Cunningham also advocates for a "proactive agenda" when it comes to the design of durable recordkeeping systems as well as interactions with records creators, stating that "we cannot take for granted ... that records ... will remain reliable, comprehensible, authentic, accessible,

[32] Ibid., 327.

[33] Adrian Cunningham. "Waiting for the Ghost Train: Strategies for Managing Electronic Personal Records Before It Is Too Late," *Archival Issues* 24 (1999): 58, 60.

[34] Ibid., 58.

[35] Ibid., 59.

[36] Ibid., 60.

and durable" without intervention.[37] While he acknowledges that pre-custodial intervention in all forms can be labour intensive, he does think that it will pay in the long run, cutting/eliminating time spent on the arrangement and description of poorly maintained records.[38] Cunningham also advocates working with software developers and vendors to encourage the "incorporation of good record-keeping functionality and self-documenting features" in their applications and products, a strategy that will not be lost on archivists who have worked with developers and programmers to plan systems.[39] The more archivists know about the software development process, the more they can do to advocate for archives-aware software development for personal recordkeeping.

Finally, Cunningham addresses the "head-in-the-sand" mentality of archivists who ignore the problem or are waiting for someone else to solve it. He implies that failing to pursue a "more active agenda" will leave us with nothing to "satisfy our researcher's need for solid, reliable, and authentic evidence of the past."[40]

In her article "The Long-Term Preservation of the Digital Heritage: The Case of Universities Institutional Repositories," Luciana Duranti proposes practical strategies based on theoretical principles for maintaining authenticity and protecting producer rights within an institutional repository. Most important in establishing authenticity are the "integrity of the environment" in which a digital entity resides and the "processes aimed to maintain them and to ensure accountability of the person or organization responsible for them."[41] This means creating a preservation methodology that allows for mechanisms that verify source, transmission, and sustainability.[42] Duranti also explains the challenging nature of institutional repositories; their "mix of documentation and data," create challenges to continued access and preservation, which is also the reason why they exist.[43]

The Need to Incorporate Archival Theory into Technology Development

An understanding of key archival theories and principles set within the electronic records context should inform the design, development, and implementation of an institutional repository. Traditional and new methods for appraisal and selection, determining and preserving authenticity and reliability, and sug-

[37] Ibid., 59.
[38] Ibid., 60.
[39] Ibid., 61.
[40] Ibid., 63.
[41] Luciana Duranti, "The Long-Term Preservation of the Digital Heritage: The Case of Universities Institutional Repositories," *Italian Journal of Library and Information Science* 1 (2010): 158.
[42] Ibid.
[43] Ibid., 159.

gestions regarding systems and skills for the modern archives and archivists to employ, all have a place within a real-life institutional repository.

Although the initial commitment can seem overwhelming – even unattainable – balancing archival theory and practice in an electronic records context will improve the chances of acceptance from all stakeholders. Additionally, archivists working within universities and larger research institutions often find their approach to managing born-digital content to be "entwined with the parent institution's perspective and strategy on digital infrastructure and institutional repositories."[44] The need to balance theory, principles, institutional perspectives, and real-life expectations is a challenge for archivists; however, it can inspire acceptance and interest from diverse campus, library, and external groups.

Keeping archival theory, institutional perspectives, and achievable goals in mind, the UNC developed a preservation-focused repository that incorporates archival concepts into its architecture. Because of institutional needs and goals, the CDR development process specifically emphasized the need to look at both the theoretical and the organizational issues surrounding digital preservation, a necessary approach given the diverse collecting streams of the CDR. The repository currently accepts born-digital special collections, digital scholarly output from the UNC community, and library-generated digitized content.

Convergent Theories in Archival Practice and the CDR

The somewhat disjointed history of the CDR makes it difficult to pinpoint a specific theoretical framework from which it was developed; it is safe to say, however, that the repository was based on "checklist" archival concepts and theory during its conception and in its development and deployment. During the first couple of years of repository discussions and planning, faculty and graduate students from the School of Information and Library Science (SILS), and practicing archivists and librarians from the university libraries played a large role in visualizing the repository that was to be developed and setting its scope.

During this time, SILS's DigCCurr Project was developed, and many DigCCurr graduate fellows were involved in one aspect or another of CDR planning and development.[45] Concepts being used to develop the Trustworthy

[44] Richard V. Szary and Erin O'Meara, "If Not Us, Who? University Archives and Campus-Based Digital Preservation Repositories," *International Council on Archives Section on University and Research Institution Archives Conference, Edmonton, AB, 15 July 2011,* p. 8.

[45] The DigCCurr Project is an initiative to develop a graduate-level curriculum to prepare students to become information professionals in the digital era. See http://www.ils.unc.edu/digccurr/ (accessed 15 July 2011). DigCCurr is in its second phase as a project.

Repositories Audit and Certification (TRAC)[46] also influenced work on the CDR. The TRAC perspective shifted focus from the authenticity of the record itself to the larger issues of audit and assessment of the viability and trustworthiness of the repository within its organizational context. Based on the CDR experience, this perspective is essential; without dedicated resources, the practicalities of repository development cannot occur.

When assessing the trustworthiness of a repository, TRAC looks beyond the preservation activities that a repository can perform; it takes into account governance and financial issues surrounding the repository:

> [I]n determining trustworthiness, one must look at the entire system in which the digital information is managed, including the organization running the repository: its governance; organizational structure and staffing; policies and procedures; financial fitness and sustainability...[47]

The Digital Curation Lifecycle model was also consulted in the conceptualization of the CDR; the model focuses on business planning and sustainability as key factors in an organization's ability to curate digital objects.[48] A lot of time and effort were dedicated to researching the system requirements for the CDR. Through the work of the Digital Curation-Institutional Repository Committee (DC-IRC), this was accomplished through feedback from diverse perspectives: TRAC, the Open Archival Information System (OAIS) Reference model,[49] and the Digital Curation Lifecycle model were all influential when deciding what technology to choose. The idea of micro-services had not yet come about, but the ideas of sustainability, flexibility, and extensibility were discussed.

During the course of repository planning and development, organizational shifts within the library changed the reporting roles of the CDR's project staff. Between the move of the project from the Carolina Digital Library and Archives (CDLA)[50] unit to the Library Systems department in 2008 and the

[46] Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist was developed to assess the trustworthiness of repositories through an external audit process. See http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed 11 July 2011).

[47] Ibid., 3.

[48] DCC Curation Lifecycle Model, http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf (accessed 11 July 2011).

[49] The Open Archival Information System (OAIS) Reference model is designed as a broad, conceptual framework for digital preservation environments. CCSDS, "Reference Model for an Open Archival Information System" (Washington, DC, 2002), http://public.ccsds.org/publications/archive/650x0b1.PDF (accessed 22 July 2011).

[50] The Carolina Digital Library and Archives (CDLA) began enhancing services to digital scholarship at UNC in 2007. The CDLA provides digital project management services and is the main digital production unit in the libraries. See http://cdla.unc.edu (accessed 22 July 2011).

dissolution of the DC-IRC, the focus on archival concepts and principles was de-emphasized; the technical development of a more traditional institutional repository became the primary concern. This shift led to decisions being made solely on repository functionality, and as a result, preservation actions that can be described as "good enough" and "just in time" were executed. It became increasingly clear that changes in organizational structure, staffing, and budget allocation could tremendously alter the perspective and trajectory of a repository project or program – losing your champion during development can threaten the viability of the project. It is, therefore, necessary to build these relationships in the earliest stages and foster them throughout the development of the project, regardless of any organizational changes that occur.

When one of the authors (O'Meara) arrived at UNC in the fall of 2009 and began work with the CDR project staff, she encouraged a shift back toward repository development rooted in archival concepts. O'Meara's perspective was developed while she undertook her Master of Archival Studies degree (MAS)[51] at the University of British Columbia. She brought to the CDR concepts that were based in historical recordkeeping, but applied in the digital environment. She also has experience with the challenges facing modern collecting repositories. One of these challenges is that the records creation, records keeping, and records preservation activities all form the chain of preservation needed throughout the life cycle; in practice, however, these records environments are usually very separate and under very little control. Collecting repositories commonly receive records from the creator at the end of a collection's active life. More often than not, procedural control was never exerted over these records. In practice, there is a de-emphasis of the need for documentary and procedural control that is stressed in theoretical descriptions of both record-making and recordkeeping environments. In modern record-creating environments,[52] registration, classification, and other procedural controls over records are unknown at the creator's level, except in highly regulated environments, such as patenting and pharmaceutical drug development. Explaining the concepts of diplomatics (i.e., persons, form, and actions)[53] as they pertain to records does not seem relevant to most technolo-

[51] While O'Meara was a student, the MAS program focused on theory surrounding *diplomatics* and research findings from the InterPARES Project. The focus on theory served as an indoctrination so that in any environment, one could rely on the core concepts from the MAS program, especially the concepts of authenticity and reliability of records.

[52] Examples of creators' environments of objects ingested into the CDR include a research lab at UNC, materials resulting from student class projects, and special collections that seek out material from personal donors.

[53] *Persons* are "the entities recognized by the juridical system as capable of having the potential for acting legally." Key persons in a record are the author, addressee, and writer. *Form* is the rules of representation that are the "evidence of the intent to convey information" and is represented both intellectually and physically within the record. *Actions* form the

gists when they are in the midst of active software development. Because of this, O'Meara decided to extract the essential meanings of the concepts so that non-archivists could relate to them: digital preservation is about maintaining the context, content, and form of a record. This distillation of preservation concepts was received positively by the project staff. During these discussions, the project team read Peter Hirtle's "Archival Authenticity in a Digital Age,"[54] an article that explains, in clear terms, the concepts of authenticity in a digital environment using illustrative examples relevant to archivists and non-archivists alike. O'Meara's work to introduce archival concepts to other project staff using clear language and concrete examples enhanced communication between archivists and technologists. In order to foster a shared, constructive, and collaborative dialogue between archivists and the communities they work with, it is necessary to communicate using common terms and concepts that everyone can understand. While the OAIS Reference model states the shared vocabulary between archivists and technologists, the nuances particular to each profession may need to be further expressed. For example, the reasons behind documenting events post-ingest goes beyond terms such as Submission Information Packages (SIP), Dissemination Information Packages (DIP), and Archival Information Packages (AIP).[55]

Both the theoretical and practical perspectives of digital preservation address the need to involve the records creator or producer during the "active phase of the life cycle" (as it is called in the OAIS model). The methodology of each perspective, however, is different. One of the core principles of the model is, "keep lifecycle stages simple, and move complexity into the functions."[56] This is somewhat different from the theoretical perspective of the intricate InterPARES 2 Project Integration Definition for Function Modeling (IDEF)[57] models, especially the Chain of Preservation Model, which visually demonstrates the complexity that the theoretical perspectives propose for the creation, management, and preservation of digital objects.[58] The records cre-

impetus for records creation and come from a will to determine a fact. See Luciana Duranti, *Diplomatics: New Uses for an Old Science* (Maryland, 1998), 83, 41, 62.

[54] Peter Hirtle, "Archival Authenticity in a Digital Age," in *Authenticity in a Digital Environment,* ed. Council on Library and Information Resources (Washington, DC, 2000), 8–23.

[55] These terms refer to packaged objects (Submission, Dissemination, and Archival Information Packages, respectively) moving through an archival repository per the CCSDS's *Reference Model for an Open Archival Information System.*

[56] Christopher A. Lee, Helen R. Tibbo, and John C. Schaefer, "Defining What Digital Curators Do and What They Need to Know: The DigCCurr Project," *Proceedings of the 2007 Conference on Digital Libraries* (2007), 49–50, available at http://www.ils.unc.edu/digccurr/jcdl2007_paper.pdf (accessed 3 August 2011).

[57] Integration Definition for Function Modeling (IDEF) is a modeling framework used for systems design.

[58] Luciana Duranti and Randy Preston, eds., *InterPARES 2 Book,* "Appendix 14: Chain of

ation and records keeping life cycle stages in the model are very detailed; the complex model provides an excellent paradigm for study, and a springboard for brainstorming system requirements.

The challenge for the CDR team was to interpret these various requirements (InterPARES Chain of Preservation Model and TRAC documentation) and develop a functional system while working with limited resources, under strict timelines, and realizing the need to serve diverse collection materials. The Chain of Preservation Model that came out of the Modeling Cross Domain of the InterPARES 2 Project provides details at all levels of the records life cycle and helps illustrate the concept of preservation used by the project team. The model is extremely complicated, but uses the modeling framework to go from the scale of the entire records environment (creation, maintenance, and preservation) to the packaged and preserved electronic records.[59] Appendices 20 and 21 of the *InterPARES 2 Book* serve as more-easily interpretable advice for records creators and records preservers.[60] Appendix 20 covers concepts that records creators would need to address for their environment. Appendix 21 spells out the preservation requirements visually represented in the Chain of Preservation Model for archivists. The model was used as one conceptual ideal in the development of the CDR.

The main area where theory was directly applied in the CDR revolved around ensuring that the digital objects would be inextricably linked to the context in which they exist within and amongst other records in the collections. We used the object model framework built into Fedora Commons repository software to maintain and communicate these relationships. The CDR has a content model that represents hierarchical relationships between objects (files and associated metadata) that was tested against various archival and non-archival use cases.

Ensuring the authenticity of a record is one of the fundamental concepts surrounding diplomatics, the science of the nature and formation of records.[61] Archivists can use the methods devised from diplomatics to test the authenticity of a record and identify alterations to it. With existing technology and limited resources, authenticity can become difficult to maintain when providing an authentic copy of a record for access.

In the future, the CDR could enhance authenticity by employing processes

Preservation Model – Diagrams and Definitions," http://interpares.org/ip2/book.cfm (accessed 3 August 2011).

[59] The IDEF model is brilliant in its complexity, but would be difficult to use as a requirements document since it involves rigorous behaviour from records creators, active records stewards, and the archives.

[60] Duranti and Preston, *InterPARES 2 Book*, http://interpares.org/ip2/book.cfm (accessed 23 July 2011).

[61] Duranti, *Diplomatics: New Uses for an Old Science*, 27–35.

similar to those in the field of digital forensics. Snapshots of operating system environments and explicit descriptions of the state of the object and its environment can be captured with these tools.[62] New methods for incorporating this information into the archival description of the records will help to ensure the authenticity and reliability of records by automated means, all within a controlled system. The CDR has the ability to do this; however, it is a matter of prioritization and resources as to what level the repository can and will integrate digital forensics tools into the pre-ingest workflow. The recent collaboration between the two fields has brought with it new techniques that might make it possible for medium-sized repositories to begin to address issues of authenticity and reliability systematically in electronic records preservation.

Before a repository can provide full preservation and curation services, it needs to demonstrate basic preservation behaviour. With the CDR, there was a decision to focus on basic, bit-level preservation. Because of this, some curation activities were not always fully addressed and for a time, this was a contentious topic within the steering committee. If the collecting scope had been narrowed and more resources had been dedicated to building the technical and ingest infrastructure, full curation services at the onset would have been more feasible. Some steering committee members felt that the curation of some of the content (e.g., the digitized surrogates of historical photographs), would not need the full curation services that a collection of born-digital personal papers would need.

This variety of content (and the diverse needs associated with it) has been a common challenge across institutional repository development projects. While the format-driven approach to defining preservation policies can be a solution for institutional repositories that aim to collect traditional content (such as textual documents that have been generated fairly recently), it can be problematic for repositories with a wider collecting focus, such as the CDR. Policies for levels of preservation and format support similar to those employed by the University of Michigan's Deep Blue[63] institutional repository is one way to delineate preservation services among various formats and content types ingested. The CDR is looking at the three main collecting streams (born-digital special collections, digital scholarly output from the UNC community, and library-generated digital content) as a way to develop tailored preservation

[62] For more information on the application of digital forensics in an archival setting, see Matthew G. Kirschenbaum, Richard Ovenden, and Gabriela Redwine, with research assistance from Rachel Donahue, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington, DC, 2010), http://www.clir.org/pubs/reports/pub149/pub149.pdf (accessed 15 July 2011). The authors provide a thorough overview of the state of the art.

[63] Deep Blue Preservation and Format Support Policy, http://deepblue.lib.umich.edu/about/deepbluepreservation.jsp (accessed 15 July 2011).

services. For example, specific preservation activities could be performed as a standard event for one collecting stream but not another.

The Carolina Digital Repository (CDR)

The CDR is an institutional repository for digital format materials produced by members of the UNC community. The main goal of the CDR is to keep UNC digital scholarly output safe and accessible for as long as needed. It also serves as a repository of historical materials that broadly support the university's academic mission. More specifically, the CDR aims to acquire UNC digital material, and ensure it is accessible, searchable, and safe from alteration. The CDR is a partnership between the UNC's university libraries, the Office of the Provost, and the School of Information and Library Science (SILS). The CDR provides a defined service to the UNC campus that directly aligns with the university library's larger role as a trusted steward of information.

Background and History

Several years before the beginning of formal repository work, research and investigation began that suggested the need for a repository for faculty research at UNC.[64] Repository development began formally at UNC in 2004 with the creation of the Digital Curation/Institutional Repository Committee (DC-IRC). A large stakeholder group with membership spanning the entire campus, the DC-IRC had an ambitious mandate.[65] The group developed a proposal that called for the funding of an institutional repository. This proposal was submitted to the Provost who allocated funding to hire programmers and begin the technical side of repository development.

In 2007, the university library took on the physical development of the repository (now called the CDR), supervising all project staff. Shortly

[64] The "Minds of Carolina" project was an initiative led by Helen Tibbo and Paul Jones, faculty of the School of Information and Library Science at UNC. The project explored how to enable faculty to self-archive or prepare their materials for deposit into a repository. See http://www.ibiblio.org/minds/innovation.html (accessed 15 July 2011).

[65] The DC-IRC's mandate was to "[d]evelop a feasible plan that will both serve the UNC-Chapel Hill's curation needs and will place the University in the forefront of such efforts in the Triangle, nationally and internationally; design a pilot institutional repository and digital preservation program in partnership with Information and Technology Services, the University Library, and the School of Information and Library Science that will support ongoing research; develop policies, procedures, and long-term digital preservation strategies to benefit the entire campus. This will include strategies to educate the campus community." Carolyn Hank, *A Progress and Recommendations Report from the Digital Curation/Institutional Repository Committee, 2005–07: Informing a Successful Institutional Repository Deployment at the University of North Carolina at Chapel Hill* (Chapel Hill, NC, 2004), 10.

after, the DC-IRC was dissolved and a smaller group formed the CDR Steering Committee.[66] This committee is composed of faculty members from SILS, representatives from library administration and the Data Intensive Cyber Environments research group (DICE), and several library staff who are directly involved in repository development. In April 2010, the library deployed a beta version of the CDR, providing access to three pilot collections through a web access portal.[67] Building such a robust and ambitious repository would certainly benefit from broader partnerships; however, they can slow the pace and make project management more complicated.

Repository Architecture and Technical Specifications

The underlying architecture of the CDR uses Fedora Commons repository software[68] connected to an iRODS data grid[69] to form a complex preservation environment. A Solr[70]-indexed, custom web access portal is used for search and retrieval.[71] The CDR employs custom METS and PREMIS profiles and uses MODS[72] as its primary descriptive metadata standard. The library collaborates with the campus Information Technology department for distributed storage and backup.

[66] The CDR Steering Committee serves in an advisory role to the project and helps determine priorities for both repository and collection development.

[67] The pilot began with collections from the Research Labs of Archaeology, the Southern Folklife Collection, and the African American Performance Art Archive. As of October 2011, there were over 48,000 objects in the repository (see https://cdr.lib.unc.edu/). Many of the collections have access controls that allow for collections that contain sensitive materials, or that have privacy constraints or copyright restrictions to be securely posted to the repository. The repository uses Shibboleth, the university authentication service, to grant authorized access to restricted collections housed in the CDR.

[68] Fedora Commons is an open-source repository tool. The Fedora object model concept helps intellectually arrange files and maintain relationships between files and metadata. See http://fedora-commons.org/ (accessed 15 July 2011).

[69] iRODS is an integrated, rule-oriented data system developed by the DICE group. The CDR deploys iRODS-based rules to automate preservation activities such as checksum verification and other file-level validation activities. See https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems (accessed 11 July 2011).

[70] Solr is an open-source, web-indexing platform that allows for faceted search and browse functionality.

[71] The CDR web access point is focused on download capabilities. Traditional IR functions such as impact rankings for articles, social media-based sharing capabilities, or faculty profiling have not been implemented.

[72] METS, PREMIS, and MODS are all established metadata schemas that allow for the standardization and management of the description of content. The CDR uses METS for packaging contents for ingest into the repository. PREMIS is used to describe preservation events that happen to objects over time. MODS is used to describe digital objects stored in the repository.

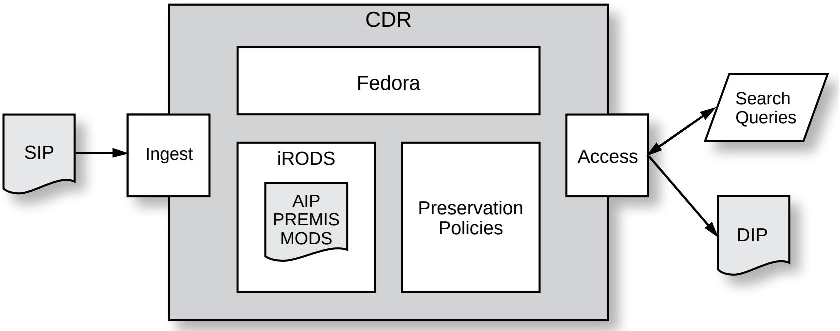

Figure 1. Technical Diagram of CDR System Architecture

Curation and Preservation Actions

The transfer of digital objects from their creator to the CDR is mediated through a repository staff member who consults with the potential depositors before accepting any content. At this time, the nature of the digital objects is discussed and an ingest survey is conducted to alert repository staff to issues concerning metadata, rights, or storage that may arise. Additionally, a deposit agreement[73] is signed before the transfer of materials occurs, ensuring that both repository staff and depositors are made aware of the rights each hold in regards to content deposited into the CDR. This pre-transfer mediation process allows repository staff to uncover foreseeable challenges to providing long-term access, and discuss these upfront with depositors in order to devise strategies for meeting both parties' preservation expectations and goals. Repository staff realized that this level of individualized service could prove to be unsustainable once the number of depositors to the CDR increases. It will certainly be necessary to move to more efficient modes of transfer, such as self-depositing and drop-boxes. This is not to say that mediation and quality control procedures will be completely abandoned, however they could be carried out in a more streamlined fashion.

Once the materials are physically or digitally transferred to the library, ingest preparation begins. First, materials are staged to an iRODS grid where they are held before metadata is linked to the objects and the collection is prepared for ingest into the CDR. Pre-ingest activities include selection at the file and object level, arrangement of files and objects, and linking user-supplied or library-generated metadata to the objects. Ingest occurs mainly through a web

[73] See Appendix A.

portal, where a METS manifest is generated through the CDR's ingest tool, Curator's Workbench.[74] This manifest is uploaded along with a brief, top-level collection description. Curator's Workbench also generates unique identifiers, checksums, and a METS profile that maps out the structure and arrangement of objects (files and associated metadata) within a collection. Upon ingest, the CDR's custom Fedora ingest service generates another checksum for each file and verifies it against the checksum from the pre-ingest METS manifest. The Fedora ingest service also generates the relationships described in the METS files within the pre-defined Fedora object model.

Currently, the CDR provides bit-level preservation of the digital objects ingested into the repository,[75] i.e., the bitstreams of the files are preserved exactly as they are deposited. In the future, the CDR hopes to move beyond bit-level preservation and integrate a suite of preservation activities such as file normalization, refreshment, and migration services. These services will be performed on select content and their deployment will ultimately depend on available financial and personnel resources.

Priorities for preservation activities within the CDR have been envisioned in a layered approach. The first layer of preservation activities, which has been instituted, is keeping the actual ingested files at the bit-level safe from alteration. The second layer of preservation activities will increase the functionality of the system so that it can perform better integrity checks upon, and monitor the overall health of, the objects held within the system. A third layer of preservation activities will include better documentation of digital collections and objects from the time they are acquired by the library to the point of ingest into the CDR. This will strengthen the chain of custody of the objects and is derived from the Jenkinsonian theoretical perspectives discussed previously. We hope to implement further preservation activities for the CDR, specifically addressing the issue of object and system integrity described by Duranti and MacNeil.

It is unclear if library staff will ever be successful in increasing the documentary and procedural controls of recordkeeping systems while the records are active and in the creator's custody. Without exerting sufficient levels of pre-custodial control over the records in these environments, will we ever have

[74] The Curator's Workbench is an open-source, pre-ingest tool developed at UNC to support the appraisal and processing of digital materials. The open source code and wiki is available on GitHub, https://github.com/UNC-Libraries/Curators-Workbench (accessed 15 October 2011). The tool is designed to improve efficiency when processing large numbers of files with custom, non-standard metadata. It assists with description, file staging, and the crosswalking of custom metadata to a standardized format.

[75] For more on bit-level preservation as compared to full preservation services, see Priscilla Caplan, "The Florida Digital Archive and DAITSS: A Working Preservation Repository Based on Format Migration," *International Journal on Digital Libraries* 6, no. 4 (July 2007): 307.

truly authentic and reliable records to ingest? Even though it is now more than fifteen years old, John McDonald's article "Managing Records in the Modern Office: Taming the Wild Frontier" rings true:

> [I]n many ways the modern office environment is not unlike the wild frontier of the last century. Instead of horses and wagons, our organizations have provided us with computers and software, telling us to charge off into the great unexplored plains of cyberspace where supposedly we can work more effectively.[76]

Pre-custodial intervention can be used to offset this issue – to an extent. Training or working with content creators earlier in the records life cycle will provide more reliable records and better metadata. Pre-custodial intervention does take extra resources from library staff, but with large streams of regular deposits from specific donors, the benefits may outweigh the costs. While we cannot fully exert documentary and procedural controls before the records come to us, we can maintain the authenticity of deposited records that are in both the library and the CDR's custody. Communicating this fact to the users of the CDR – who may not realize that the content they are accessing was not created by the library – is a challenge that remains to be addressed.

Challenges

Even with stakeholder support and dedicated staffing, there were numerous challenges that tested the success of the repository and how it would fulfill its mission. In the fall of 2009, the University Librarian gave project staff a deadline to have a working repository with a small set of pilot collections ingested by April 2010. This tight deadline inhibited the realization of certain preservation activities within the CDR for the soft launch of the project. Focus was placed on how to ensure that core repository functions, such as basic ingest, storage, display, download, replication, and disaster recovery, would be carried out. To move forward with the development and launch of the repository and meet this deadline, repository staff members made compromises. Staff decided to prioritize basic functionality across the repository and build further enhanced services in collecting, preservation, and access over time.

Traditional archives have a clear collecting mission, scope, and audience. With institutional repositories, however, collecting areas, scope, and stakeholders are sometimes blurred, ill-defined, or not defined at all. There is an identified need for stewardship of digital scholarship within the UNC community, both within the library and externally from faculty and students, and the CDR is trying to hone in on its prioritized collecting tracks based on these

[76] John McDonald, "Managing Records in the Modern Office: Taming the Wild Frontier," *Archivaria* 39 (Spring 1995): 71.

needs.[77] As more objects are ingested into the repository, CDR governance will need to decide how to narrow the collecting scope. By focusing on selected content areas or formats, the repository can demonstrate value and relevance to users and other stakeholders. Having a broad collecting scope means that not everything that is ingested into the repository fulfills the definition of a record, even when a broader definition is being used. As the repository matures, so will the collecting areas. These areas will help determine the levels of preservation needed for each category.

Both the theoretical and organizational approaches support the idea that archivists should have a proactive and involved role in the development of repositories, including policy and technology decision-making. However, with only one archivist on the project staff and a few more in the Steering Committee, it is highly likely that decisions affecting the preservation framework will be made with little input from archivists. Although the CDR has built in mechanisms for consultation by archivists before large technical decisions are implemented, separate departmental affiliations and physical locations of archivists and technologists encumbers this process. While large institutions can have the advantage of having more resources, archivists can sometimes get lost in the organizational structure where they will not be able to make an impact. Access to technology and clear communication channels to technologists is the key to making sure that this does not happen. The CDR has been a great example of the benefits of keeping organization-wide dialogue between technologists and archivists clear and open.

Conclusion

The CDR's current priorities include enhancing existing preservation activities and using a measurable standard to assess performance.[78] Repository staff members continue to build on existing preservation activities within the repository to demonstrate the integrity of the system itself, and ensure the reliability and authenticity of the records it preserves.

In order to remain true to themselves and the profession while continuing to move forward and stay relevant, it is necessary for archivists to reflect on archival theory while remaining open to practical innovations. This is just one of the many challenges archivists face when tasked with designing and implementing institutional repositories. The case of the CDR reflects on this

[77] Collecting tracks currently include faculty and student scholarship such as datasets, digital research materials, and published material that the UNC community deposits into the institutional repository; institutional electronic records that University Archives collects; born-digital special collections that are transferred to Wilson Library; and digitized special collections within Wilson Library.

[78] For example, using TRAC as a guideline to form an internal framework for assessment.

challenge and others, including finding balance between finite institutional resources and conceptual ideals presented in archival theory and research.

Returning to our original question, *Is it possible to design and implement a preservation repository founded on traditional archival theory and principles*? Yes, to an extent, and depending on repository aims and institutional perspectives. It is possible to reach some goals, and it is necessary to compromise on others. The CDR is an example of these goals and compromises realized.

Compromise was necessary in three areas. First, with respect to preservation activities, we found that digitized content does not need the same preservation environment as electronic records, and consequently chose to use a layered approach to preservation activities. Second, as far as descriptive practices were concerned, we chose not to use Encoded Archival Description (EAD) within the repository, but where appropriate to link EAD finding aids to born-digital collection objects that are in the repository. Third, we chose a fairly large and undefined collecting scope over a more thorough, formal, collecting policy.

Unexpectedly, we found that this large collecting scope did have its benefits. For example the CDR was able to obtain more funding opportunities than a smaller repository, such as one for just born-digital objects, would have been able to secure. Additionally, because of the diverse nature of the CDR's collections, it received more recognition within the library and the professional community.

In conclusion, we offer some thoughts on how to sustain a preservation repository based on our experience with the CDR. Start, and continue, to build the repository in staged layers. If you currently cannot build an actual repository, begin to think about future requirements, and in the meantime build a storage space that incorporates bit-level validation and other basic preservation activities that can be performed over a file system. Base your repository architecture on theory and best practices, but do not strive for perfection or unattainable goals. Acknowledge, and be able to communicate, that a repository development project requires a serious investment of resources. Stress to administration and project staff that archivists need to be there when policy and technical decisions are being made. Work to get more library staff engaged with the repository by making it part of their daily workflow. Finally, balancing preservation and user needs with a shrinking budget is challenging, but demonstrating use and value can help build a case for the ongoing commitment of resources.

Appendix A: Carolina Digital Repository Non-Exclusive Deposit Agreement

Please take a moment to read through the terms of this agreement. Your signature is required for the University Library to reproduce, translate, and make your submission publicly available through the Carolina Digital Repository.

By signing this agreement, you (the author(s) and copyright owner(s)) grant the University of North Carolina at Chapel Hill (UNC) the non-exclusive right to reproduce your submission, translate the submission to any medium or format for the purpose of preservation and public access, and/or publicly and globally distribute your submission in electronic format.

You also grant that UNC may keep more than one copy of this submission for purposes of security, back-up, and preservation.

You agree that the submission is your original work, and/or that you have the right to grant the rights contained in this agreement. You also agree that your submission does not, to the best of your knowledge, infringe upon anyone's copyright.

If the submission contains material for which you do not hold copyright, you agree that you have obtained the unrestricted permission of the copyright owner to grant UNC the rights required by this agreement, and that such third-party owned material is clearly identified and acknowledged within the text or content of the submission.

If the submission is based upon work sponsored or supported by an agency or organization other than UNC, you agree that you have fulfilled any right of review or other obligations required by such contract or agreement.

UNC will not make any alteration, other than as allowed by this agreement, to your submission.

□ I ACCEPT the terms of this non-exclusive deposit agreement

□ I DO NOT ACCEPT the terms of this non-exclusive deposit agreement

Name (Please print):____

Signature: ____

Date: ____