PROCEEDINGS OF

DigCCurr2009

say : dij-seeker

DIGITAL CURATION

PRACTICE, PROMISE & PROSPECTS

Edited by:
Helen R. Tibbo
Carolyn Hank
Christopher A. Lee
Rachael Clemens

April 1-3, 2009
University of North Carolina at Chapel Hill, NC USA

INSTITUTE of Museum and Library SERVICES

cni Coalition for Networked Information

National Archives and Records Administration (NARA)

UNC UNIVERSITY LIBRARY

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Proceedings of DigCCurr2009
Digital Curation: Practice, Promise and Prospects

This volume contains contributions to DigCCurr2009, held April 1-3, 2009 at the University of North Carolina at Chapel Hill.

ISBN 978-0-578-02215-4

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

School of Information and Library Science
University of North Carolina at Chapel Hill
CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
Phone: 919.962.8366
Email: info@ils.unc.edu

The Survival of Records (and Records Management) in the Twenty-First Century

Kenneth Thibodeau
National Archives and Records Administration
Washington DC, 20740 USA
Ken.Thibodeau@nara.gov

ABSTRACT

This paper discuss the nature and role of records and records management in relation to the contemporary cyber-landscape, and describes how the principles of records management inform the retention of electronic records and what changes in perspective and method are needed given the digital domain.

Keywords

Electronic records, recordkeeping, digital preservation, InterPARES.

1. INTRODUCTION

The first thing that might come to mind, on hearing the topic of the survival of records in the twenty-first century, is the problem of digital preservation. Obviously, the preservation of records in digital form is crucial, but while preservation is necessary, it is not sufficient. Given the open-ended proliferation of new forms for creating, capturing, and combining information in digital systems, of ever wider, more diverse, and more highly specified ways of applying digital technology in the conduct of business, of the expanding capacity to re-use and even re-purpose digital data to satisfy a variety of both planned and spontaneous needs, it is also necessary that records continue to be recognized not only as a distinct class of information assets, but also as one which merits special attention. It is not sufficient for records keeping to be seen as a necessary part of doing business. Necessity of this sort is all too often the daughter of laws, regulations, and other external requirements. Records that are kept as the result of such external forces are easily relegated to the sidelines and perceived as having marginal value in the accomplishment of practical objectives, in strong contrast to the immediate value of real time online transactional processing, multidimensional analytic processing, geo-referencing, experiential computing, and other powerful tools made available by the application of computer science and engineering. If records slide to the margins of the conduct of business, records management will diminish with them. To survive and prosper in the twenty-first century, records and methods for managing them need to have a vibrant, organic relationship to the conduct of affairs. To be a vital contributor to corporate, institutional, individual and societal success in the twenty-first century, records management must deal cogently and comprehensively with the increasing permeation of digital information in the conduct of affairs.

Cyberspace is so different from what has gone before that it has been characterized as the fifth dimension, one which enables us to create and do things which are not possible in the space-time continuum.[1] The creation, capture, and communication of information in cyberspace is drastically different than in four dimensions, because the digital dimension breaks the bounds of space and time that constrain the information technology of hard copy documents. To face the challenges and take advantage of the possibilities offered by digital technology, we need a richer and deeper understanding of the nature of digitally encoded information and how such information can be and can be managed as records. We need to be better able to apply the knowledge we have of records and records management in the digital realm; translating it in terms that make it effectively operative in cyberspace; adapting it where necessary; and also abandoning those concepts and techniques that are not viable in cyberspace. If we fail to do so, we run the risk of seeing records management become an increasingly esoteric exercise, divorced and isolated from the mainstream of affairs.

The first thing records managers need to do is to acknowledge that established knowledge and methods have limited applicability in the digital dimension. The second thing – logically but not necessarily chronologically – is to recognize the opportunities that digital technology creates for managing records in ways that might exceed by far anything possible in four dimensions. Consider just two basic topics: What are electronic records, and how should we organize them? For both topics, let us consider how well or how far established knowledge applies in the digital realm; some obvious limits to its applicability, and alternatives that the digital dimension opens up.

2. WHAT ARE ELECTONIC RECORDS

Conceptually, electronic records are not radically different than traditional records. Traditional definitions and concepts identify fundamental characteristics that are independent of how records are constructed, encoded or stored. A record:

- is a unit of information;
- is made or acquired in the course of activity;

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

1 Christopher Greer. 2007.

- contains evidence or information about that activity;
- is kept for use in subsequent activity or for reference, and
- is organically related to other records of the same records creator and activity.[2]

But problems arise when we try to deal with electronic records empirically. From an empirical perspective, we are not yet able to characterize them in a way that would enable us to articulate surefire methods and processes to maximize their value in the conduct of affairs. Records as we know them are artifacts of a particular genre of information technology, the technology of hard copy, where information is affixed to a physical medium in a hard and fast way. Digital encoding fundamentally alters the relationship between medium and message. In the digital realm for all practical purposes the relationship between medium and message is immaterial. Digital data move frequently and repeatedly from one physical medium to another: from silicon chips to magnetic drives, solid state memory, optical discs, copper cables, optical fibers, and electromagnetic waves, but the information objects the data constitute persist unchanged across such physical transformations.

The message here is not the medium, and it is certainly not about the durability of the medium. In the realm of hard copy, people could, did, and do use both permanent and short-lived media: from the clay tablets of the Babylonians to the wax tablets of the Romans, from cheap newsprint to archival quality paper. The same is true in the digital realm. There is no physical or chemical barrier to permanent digital media. You could write digital data on clay tablets. It's just that be business case for doing so is lousy.

The basic difference in the relationship of electronic records to their physical carriers is but one of many ways that electronic records substantially differ from traditional records. Additional differences will emerge as digital technology evolves. Menne-Haritz suggests that the shift from 'written' to electronic communications is as epochal as the shift from oral to written communications in past millennia. Articulating a variation on the theme that form follows function, Menne-Haritz points out that changing forms enables and impels changes in functions and that, conversely imposing old forms may constrain our ability to function effectively [1]. New biological species emerge through the gradual accumulation of mutations operating on a very small scale, but widespread speciation, as well as its evil twin, extinction, is often driven by large scale disruptions, such as global changes in environmental conditions. We should expect that in the twenty-first century some older forms of records and ways of managing them will not survive. They will either become extinct or retreat into ecological niches in the information landscape. We should also recognize that both marginally different and radically new types of records have already emerged in cyberspace, including new genres that have no parallel in the world of paper and other hard-copy records. With types of electronic records which appear to be counterparts of familiar hard-copy documents, we must be able to recognize and respond to even minor mutations that either imperil or promise to improve the value of records. New species of records will continue to emerge apace with the evolution of digital technology.

A basic difficulty we encounter in trying to apply knowledge derived from experience with traditional records in cyberspace is how to identify an individual record, an item of information that cannot be further decomposed without loss of 'recordness.' In physical space, unit records are often congruent with the physical media on which they are inscribed: a piece of paper, or several pieces stapled or clipped together, or a roll of motion picture film. In cyberspace, what appears as a single document may actually consist of data that are stored in numerous separate objects, each with a different structure and semantics; assembled by means of an intermediate object, such as a view on a database; organized according to the specifications of a form; and presented according to the dictates of one or more style sheets. There is nothing – no single object – stored in the computer system that corresponds to the document presented to a human in such a case. Such a document – which I call a "pseudo-document" or "pseudoc" for short – fails to satisfy one of the defining characteristics of a record: it is not kept.

How then do we deal with the case where a pseudoc is the exact counterpart of a traditional record and serves the purposes the traditional record would have served? The first InterPARES project studied numerous cases of this sort and concluded that it is literally not possible to preserve such records in electronic form. It is only possible to preserve the ability to produce copies of the records. This led to the articulation of sets of benchmark and baseline standards for maintaining the ability to produce authentic copies of such records [2]. But, if we are not preserving the records, but only the ability to reproduce them, what is the stuff we do keep in digital form? Analyzing cases which encompassed both digital replicas of hard copy records and electronic records which have no traditional counterparts, Duranti and I arrived at the necessity of distinguishing two different classes of records in such cases: the information kept in digital form, which we designated as a 'stored record' and the rendering of this information in a form suitable for human use, characterized as a 'manifested record' [3].

Keeping a manifested record in the manifested form would be redundant given the stored record. Thus, the term, 'manifested record,' is shorthand for "a copy of the record we would have kept had we decided to keep it in human readable, rather than digital, form." The manifested record may exactly reproduce the content and form that a human author saw when the record was created, but it may also be a document that a computer application produces de novo from data extracted or derived from the contents of one or more other documents created either by human authors, by processing of externally originated data, or by system to system interactions. The output from one or more stored records is not necessarily a record. It may be a temporary, evanescent display which is not saved. Nevertheless, given proper procedures and controls, a manifested record can be an authentic copy of a record.

How do we identify and delineate individual digitally stored records? In simple cases there can be a one-to-one correspondence between a digital file and a manifested record. But in many cases, the relationship between stored and manifested records can be many-to-one, one-to-many, and even

2 See, for example the definitions in the ISO Records Management Standard, ARMA, the InterPARES project.

many-to-many. The diversity and complexity of ways that pieces of information can be stored digitally and combined and processed either to generate a human-readable manifestation or for uses in system-to-system transactions led Batley to argue that in cyberspace we should concentrate on "information chunks." Psychology defines an information chunk as a unit of perception and meaning, or a digestible unit of information. Batley prefers the definition of information chunks within documents offered by information architecture gurus, Rosenfeld and Morville: "The most finely grained portion of content that merits or requires individual treatment" [4]. Batley asserts that "Each chunk will have a status and a function (and a form) that will merit its definition as a discrete record. Clearly then, it is not the document itself that merits the status of a record, rather it is the chunks of information content each document contains that are the records that need to be managed' [5].

This approach, however, fails another of the defining characteristics of a record: its relationship to an action or activity. Some information chunks can be used in many records of many different actions. For example, data identifying a customer will appear in every interaction with that customer. They will appear even in multiple records related to a single transaction, such as the order, shipping manifest, and invoice for a sale of merchandise. While the data identifying the customer are necessary parts of the contents of such records, the chunk itself carries no clue of the actions or records in which it participates. Batley's argument that managing electronic records entails managing information chunks aligns with a way computers commonly identify, store and process information, but the concept of raw information chunks needs to be supplemented in order to provide even a minimal basis for either providing adequate information about past activity or contributing to subsequent activity. The InterPARES project distinguished information chunks that occur in or contribute to the reproduction of manifested records, calling them "digital components" of electronic records. A digital component is a bit string that is necessary to produce a manifested record. A digital component might be called a "record chunk," but it is below the level of a record. To rise to the level of a record, an information object must have a defined, or at least definable, relationship to activity. We need to manage at the level of digital components, as well as at the level of records, in order to ensure that we can reproduce manifested records from their digital components.

Digital components should not be limited to portions of content, as they are in the Rosenfeld/Morville definition. A digital component might not consist of content, but define what content should be included in a manifested record, such as in a database view. It could also specify the semantics, syntax, or presentation of the record; for example, a statement in an XML Schema, an Xpath query, or a Cascading Style Sheet. Basic objects, such as the dynamic load library for a typeface or a color space should be treated as digital components if they are necessary for the output of an authentic copy of a record. Overall, there are four categories of digital components of records: (1) composition data, which tell the system what form and content data belong to a document, (2) the content data, (3) the form data, which determine how the content is arranged and presented, and (4) rules. Several different types of rules can shape the production of a manifested record. For example, rules may define the conditions or circumstances in which the record can be reproduced, or they may exclude or include certain elements of content, depending on a user's privileges, or they may define links or hyperlinks between documents or parts of a document [3]. Again, they should be retained as long as any manifested record which they control is needed and they should be managed to ensure that authentic copies can be produced.

But we are still left with the question of how to distinguish individual stored electronic records in cases where there is not a one-to-one correspondence between a stored and a manifested record. A record is something that is kept. Therefore, a stored electronic record must be a persistent object that is maintained in a computer system. A record provides information or evidence about an activity or a state of affairs at a particular time; therefore, the persistent object must contain fixed, invariant data. But it is not necessarily the case that the stored record itself provides complete information about one or more actions or a particular state of affairs. This is not peculiar to the digital realm. At least since World War II, it has been extremely common that the complete 'record' of a single action is contained in many different records.[3] Furthermore, a stored electronic record does not necessarily provide such information or evidence directly in the form in which it is stored. There is no a priori reason why an actor could not create records that require some combination with other records, or some processing in order to deliver meaningful information about past actions or situations. Again, this is not peculiar to cyberspace. Many governments, for example, do not keep birth certificates as distinct documents. Rather they capture the necessary data about each birth in a registry that contains the same data elements about all births within the jurisdiction. An individual's birth record can be produced on demand as a separate document by copying the relevant data from the registry onto a blank birth certificate. This is essentially analogous to the production of a manifested record from a stored digital record, with notable differences that computers provide much more flexibility in how the data about individual activities or situations are recorded and – because they not only capture, organize, and store data but also participate in the execution of business processes – can more reliably enforce data quality by embedding business rules in the execution of processes. Thus, a stored electronic record can satisfy the requirement that records provide information or evidence if it can be used to produce one or more manifested records that directly communicate such information or evidence.

We can, then, formulate a simple set of three criteria for identifying a stored electronic record: it must be (1) a persistent digital object that (2) contains fixed information about an activity or the state of affairs at the time when the action was done, and that (3) can be used to produce one or more manifested records.[4] The stored record may be a single information chunk or digital component, but it might just as well contain many thousands of such elements. There is no a priori limit to the structure or content of a stored electronic record and it may contain data in any one or more of the four types of digital components; that is, content, composition, form, or rule.

---

3 Gerard Naud and Christine Naud. Gazette des Archives.

4 The third criterion should be expanded to include records which are not manifested, but produced from stored records for interactions between systems where no human is involved.

3. HOW SHOULD WE ORGANIZE ELECTRONIC RECORDS

The properties of records inevitably impact methods for managing them. Methods that we have come to think of as fundamental to managing records can be seen on reflection also to be artifacts of hard copy technology. Traditional filing systems are based on the physicality of hard copy records, on location or more specifically on collocation in file folders, filing cabinets and file stations. It is certainly possible to import such approaches into cyberspace, at least iconically, as we see in Windows file management and email management products. This approach is embodied in records management applications that implement the Department of Defense standard, DoD 5015.2-STD. These applications effectively create virtual filing cabinets in cyberspace and allow us to manage electronic records as if they were in metal cabinets. But possible does not equate to optimal. What is important in the digital realm is neither the media on which information is recorded, nor the physical place where it is stored, but the possibilities that digital technology creates both for the forms of information that can be created and the ways they can be organized and used. The diversity of ways in which data can be organized; for example, in relational and object-oriented databases, data warehouses, and geographic information systems, are obvious advantages of digital technology, as are the possibilities for multiple simultaneous arrangements and for virtually instantaneous recombinations.

Do records exist in such applications? Can they exist? We know that it is certainly possible to produce manifested records from such applications. Selected content of databases is commonly output in the equivalents of traditional forms and reports. But the output of a system is not the same as what it contains. If so, how can we manage the records? One possibility might be to take them out of such applications and put them into virtual filing systems. But to do would diminish their usefulness in the conduct of business for the basic reason that they would exist apart from the systems used to execute business processes and most likely in formats that would not be useful in these processing. Breaking the organic links between records and activities would also diminish their value as records. If you want to know what an organization did and you had two places to find out: one the system the organization used to conduct its business and the other a repository where it put special forms of information that satisfy some abstract criteria for evidence, which would be the better source? Other things being equal, the system that contained the information used in the conduct of affairs, in the forms in which it was used, would be the better source. But isn't the fundamental purpose of creating and keeping records to provide a privileged source of information about prior activity? Does this represent a dilemma? How can we manage records if we don't put them in filing systems? There are other ways, made possible by digital technology. One method that is emerging is called Records Management Services. It was initiated by the National Archives and Records Administration, articulated in collaboration with 19 other federal agencies and is now being developed, with much broader participation, as a voluntary standard under the aegis of the Object Management Group. In brief, Records Management Services provide methods for identifying records in practically any type of computer application and form managing them within their native applications [6].

4. THE ROAD TO PROSPERITY AND SURVIVAL IN CYBERSPACE

Can we discern a path for records management that will enable it to survive and prosper in cyberspace? Certainly, the basic role of helping an organization to determine what records it needs to create and keep remains essential in cyberspace, with some adaptation, such as determining what records the creator needs to be able to manifest, rather than to keep. While records managers cannot personally have sufficient IT expertise even to implement decisions about what records to keep, they must be able to work with and guide a range of IT specialists in developing and maintaining systems that meet requirements for records creation and retention. But activities like this essentially translate traditional activities into the digital realm. That is not likely to be sufficient for records management to truly prosper.

Records management can contribute to realizing the potential value of digital technology in a way that best satisfies the needs of organizations and individuals in the conduct of their affairs. To do so it must be able to identify and show how records exist in business systems and it must offer methods for managing electronic records that convincingly add value in the conduct of affairs. For this, the discipline must move beyond established knowledge and methods.

Not only the specific knowledge, but even the kinds of knowledge that have stood us in good stead in the hard copy world cannot migrate into cyberspace. We should not seek to develop the digital equivalents to the kinds of expertise we had, for example, in the relative merits and drawbacks of end digit filing, or the differences between diazo and silver halide films. For one thing, the digital realm is too big and complex for us to develop sufficient knowledge about the different ways information can be encoded, organized and used. For another, by the time anyone could develop such mastery of any particular digital data type, the technology would have changed.

Information and networking technologies have substantially raised the bar on expectations for information used in the conduct of affairs, and they will raise it even higher in the future. Both competition in the market place and improvements in computer science and engineering expand the range and the amount of things digital technology can do for us, increase the flexibility and speed with which they can be done, reduce associated costs, and even make it possible to do things that are impossible outside of cyberspace. Hallmarks of raised expectations include availability, ubiquity, fluidity and responsiveness. Availability: the information needed at any time should be easy to find and access. Google is the norm and the browser is increasingly the form. Ubiquity: given the Internet, intranets, wireless, and the convergence of voice and data communications, information should be available wherever it is needed, regardless of where it is stored. Fluidity: it should be easy to change both the content and the form in which information is presented in response to different needs. This includes both adapting data to the device on which it is delivered, from a desktop monitor to a hand held PDA to a telephone, and the ability of users to select among different data types; such as, “Click to enlarge” images and tabs to select either product specifications or customer reviews in an online sales application. Responsiveness: information should be suited to business needs regardless of how complex, convoluted, demanding or dynamic they are. We already rely on digital data

to build, operate, and maintain aircraft carriers; to identify specific genes that make certain individuals more susceptible to particular diseases; to find and guide us to a good restaurant in a city we are visiting for the first time; and not only to conduct wind tunnel experiments, but even to modify those experiments in real time based on the performance of different sensors. The range, richness and usefulness of digital technologies are likely to increase.

What, then, is the potential for records management to add substantial and readily recognized value in this context? Just as the basic definition of 'record' does not need to change in cyberspace, the basic knowledge that records managers have of the most important information assets an organization has, its records, is a solid foundation for helping the organization transition to and operate in cyberspace. After all, records managers are not alone as neophytes in the digital dimension. Indeed, records managers are advantageously positioned to help organizations navigate in cyberspace because of the simultaneously broad and deep insight they have into what types of information are used in what parts of an organization and for what purposes. They can use this advantage to promote one of the basic objectives of records management: to ensure that all of the right information and only the right information is delivered to the right person at the right time to meet organizational needs.

Optimizing the utility of information entails the superposition of three architectures: information architecture, system architecture, and social architecture [7]. Information architecture is the organization of an information space, including the information it contains, to facilitate discovery and delivery of required information. System architecture is the organization of information technology components to provide functional, data and other support for specified requirements Social architecture is the organization of social networking and other social media tools and user experience capabilities to optimize the usefulness and use of digital information. Records management knowledge is most closely aligned with information architecture. It is not of much use in building or sustaining systems architecture, and effectiveness in using social architecture tools requires a very different type of expertise. However, records managers can leverage their overarching knowledge of what types of information are used in different parts and processes of an organization to contribute to a coherent approach to optimizing information assets, one that spans all three architectural domains.

5. REFERENCES

[1] Angelica Menne-Hartiz. Business Processes. An Archival Science Approach to Collaborative Decision Making, Records, and Knowledge Management. Kluwer Academic Publications. 2004.

[2] The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project. http://www.interpares.org/book/index.cfm

[3] Luciana Duranti and Kenneth Thibodeau. The Concept of Record in Interactive, Experiential and Dynamic Environments: the View of InterPARES. Archival Science (2006) 6:13–68.

[4] Louis Rosenfeld and Peter Morville. Information Architecture for the World Wide Web, Second Edition. Designing Large-Scale Web Sites. O'Reilly. August 2002.

[5] Sue Batley. The I in information architecture: the challenge of content management. Aslib Proceedings. Bradford: 2007. Vol. 59, Issue 2, p. 139-151.

[6] Object Management Group. Government Domain Task Force. Model Driven Architecture Models for Records Management Services. http://gov.omg.org/gov-rfp-rms.htm

[7] John Mark Ockerbloom. Promoting Discovery And Use Of Repository Content, an Architectural Perspective. Partnerships in Innovation II Symposium. College Park, MD. October 8, 2008.