N.C. Department of Cultural Resources

Archival Process for Data and Image Preservation: *The Management and Preservation of Digital Media*

April 1, 2008

Best Practices for the Long-Term Retention of Electronic or Digital Records

This Best Practices document serves to provide guidance to both the creators of electronic records and the custodians of trusted digital repositories, including information technology support staff, who wish to maintain the information electronically over time.

I. The Management and Preservation of Digital Media: An Overview

Digital records have taken over many of the functions that paper records once served. Like their older counterparts, digital records contain evidence of government responsibilities, citizen rights, public and private economic activities and financial transactions/obligations, scientific projects, and historical events and trends. The volume, complexity, and pace of the advances in digital media themselves, however, require the careful and consistent management of digital records if accountability and the preservation of digital records are to be assured. The integrity and accessibility of digital records also rest upon planning, documentation, and committed custodianship throughout their life cycle to an even greater degree than with paper records. Digital information is especially vulnerable to changes in software and hardware. Digital storage media, especially access technologies, are also subject to deterioration. Like every other medium or recording technology, digital technology is open to error, misuse, or fraud. In brief, to be available today, tomorrow, and into the next century, digital records must have both proper management and long-term (and in some cases, permanent) preservation. For digital records that are deemed permanent or archival, their durability needs to approach that of microfilm.

To help assure the security and preservation of records with enduring historical, evidential, or legal value, especially in the event of a human-made or natural disaster, microfilm is preferable because it is not dependent upon complex technology. Properly processed and housed microfilm lasts for hundreds of years and can be read with a magnifying lens and light. Microfilm is also an acceptable medium as evidence according to G.S. § 8-45.1 (a). It should also be noted that G.S. § 8-45.1 (b) and G.S. § 153A-436 (f) specifically prohibit the use of "computer-readable storage media" for "preservation duplicates . . . or for the preservation of permanently valuable records . . . except to the extent expressly approved by the Department of Cultural Resources . . . ." (See the texts for G.S. § 8-45.1 (a) and (b) and G.S. § 153A-436 in "Best Practices---Legal Admissibility Standards" below.)

Many public agencies and corporate organizations remain ignorant or not fully aware of the complexities of dealing with digital records. Often organizations devote greater effort to creating or receiving digital data than to its long-term maintenance and management. Managing digital records and information adequately, maintaining their authenticity, and assuring their legal acceptability all require an infrastructure containing certain detailed elements. These include policies and procedures; planning; trained staff (including assignment of specific responsibilities for data management to specific staff members, such as digital data archivists or managers, trained for their roles); and physical systems and facilities, including a digital repository.

While there is as yet no viable long-term strategy to ensure that digital information will be readable in the future, one methodology likely to prove valuable in helping assure future accessibility and legal admissibility of digital records is to gain control of and preserve information about digital objects and to manage this information in a formal, electronic record-keeping system for collections of digital objects. Currently the Department of Defense standard, DOD 5015.2-STD provides specific requirements for software applications that manage digital records. Other similar standards remain in early development. Other choices for digital record-keeping systems include digital repositories and digital asset management. Consideration needs to be given as well to the use of open source software and open architectures.

The creation and management of a digital repository---whether on an institutional scale or as a local digital storage server, component, or similar device---require attention (detailed further below) to six broad areas or functions. These functions are also a critical tool for ensuring the future legal admissibility of an electronic record:

- **Ingest**, or acceptance of the data or information and its preparation for inclusion in the repository;
- **Storage**, or long-term storage and maintenance of the data with appropriate procedures for preservation and error-checking;
- **Data management**, or maintenance of databases of descriptive metadata, appropriately updated and preserved;
- **Preservation planning**, including updating policies and procedures and monitoring the external environment, including the development of new technologies;
- **Access**, or management of the means by which users find, ask for, and receive data;
- **Daily administration**, including interaction with users, problem-solving, negotiation with data donors, and overseeing performance of the system.

(These functions are based on the Reference Model for Open Archival Information Systems [OAIS]. For specifications concerning this model, please consult the Electronic Records Archivist, Government Records Branch.)

State and local government offices accept and produce many different types of digital documents. In some instances, offices scan paper and create an image which then becomes the record. In others, offices accept electronic documents as part of their workflow and this document becomes the record. Additionally, state and local entities also create and manage large data files and databases such as geospatial information. This document is intended to address information both created and accepted by entities as a part of their business practice.

Before acceptance, data should be inspected and verified for operational use as the source intended, and for authenticity, integrity, and freedom from computer viruses. Restrictions or other conditions involving confidentiality or privacy, as well as proper retention and disposition provisions, need to be established. Data integrity must also be established through message digests or signatures, assuring that the data itself, its documentation, and all other descriptive and packaging information agree with that provided by the source. Digital validation should follow establishment of the data's integrity. The identity and integrity of the data must be periodically and systematically verified through such mechanisms as the Secure Hash Standard

(SHS) and Secure Hash Algorithms (SHA), the designated standard of the National Institute of Standards and Technology (NIST). Long-term preservation and use of digital data also depend upon the preservation of metadata and data documentation.

Organizations must also assemble methodologies, systems of hardware and software, and physical facilities to record, access, document, and protect digital data. Digital media themselves must be regularly tested and sampled for deterioration and continued accessibility. Provision must also be made for conversion or migration to new formats, storage media, and technologies. A digital risk management plan may include regularly scheduled migration of archival digital objects to new media. Care must be taken that hardware and software are maintained that can migrate archival data to new media. Documentation must be created and maintained that records information about all data formats, each type of media, required environmental conditions, processes for maintaining archival characteristics, and efforts to reduce risk. The digital data archivist or manager or a team of specialists should also assess data formats as digital technology advances and plan for formats that will become obsolete. Digital data will not be readable, useable, or legally acceptable, in the future without active management in this as in every other function listed above.

With regard to data format, documents need to be associated with useable data to assure sustainability and access. The capability of formatting the data contained in the document, or carried with the document, is important. Extensible Markup Language (XML) is a widely used and often preferred method for providing access to the data contained in the document. Along with industry standard definitions for the data, XML can provide both a standardized common dictionary and a common data structure for records custodians.

When considering XML, it is important to remember that agencies often do not create the documents submitted to them for recording, but rather accept and index documents submitted by others. Therefore, it is essential they adopt compatible standards in order to reuse what already exists. For example, the Property Records Industry Association (PRIA) has XML standards for county recorders. Standardization in the area of data formats will continue to be an issue requiring the closest attention by all stakeholders as well as interested members of the private sector.

Detailed written policies are needed for both active and long-term data management; records retention and disposition; appraisal and preservation; and disaster preparedness and recovery. Policies and procedures should address issues of confidentiality and privacy. They must be reviewed periodically and audited regarding enforcement and compliance.

Finally, physical maintenance of digital records requires stable, secure, environmentally controlled storage and operational facilities within a larger framework. This framework should include offsite facilities for storing duplicate copies of digital media as well as vital records including microfilm copies of vital records kept in paper format, and system backup copies that will be available after a natural or human-made disaster.

Best Practices for Archiving Electronic Records:

- Maintain at least 3 to 4 copies of the record. One copy should be designated as the preservation master; one copy should be designated as the access record; and one record should be designated as back-up. Having four copies allows margin should one copy fail. At least one of the duplicate copies should be stored off-site to ensure the information is preserved should an unforeseen disaster occur.
- Provide bit-level preservation storage of the record. If the preservation strategy includes migration of data, keep original bits for future solutions. Bit-level preservation includes maintaining the environmental controls to ensure optimal survival.
- Work from a copy of the material when migrating or making changes. Information may be lost during migration. If you work with the original copy, the information may be permanently lost.
- Metadata, secure hashing algorithms (SHA), and checksums as well as the data must be maintained and bundled together in order to preserve the integrity and admissibility of the data.

Best Practices—Policies and Procedures:

- Create and update policies and procedures defining proper development, maintenance, and use of the system. They should be available in electronic and hard copy print formats. These policies and procedures should include the metadata file required to interpret the records as well as technical components and characteristics necessary for reading, processing, accessing, using, and processing of records.
- Hold periodic training, regular retraining, and support programs that insure staff understands the policies and procedures.
- Update documentation about all permanent or archival electronic records sufficient to specify all technical characteristics necessary for reading and processing the records; identify all defined inputs and outputs from the system; define the contents of the files and records; determine restrictions on access and use; and understand the purposes and functions of the system.
- Describe update cycles or conditions and rules for adding information to the system, changing information in the system, or deleting information.
- Establish a security back-up routine based on best practices (e.g. daily, weekly, and monthly or as frequently as needed) to protect the information assets. Back-up materials should be stored off-site in case data restoration is needed.
- Establish secure off-site storage for all system password and operating procedure manuals (e.g. a bank safety deposit box.)
- Offices should have a robust disaster preparedness plan in place which addresses all copies of the data as well as identified off-site storage sites. Entities should identify critical series that would be needed to open the office should a disaster strike. Additionally, the disaster plan could have a mixed strategy of both warm and hot site recovery. Hot site recovery sites mirror all information at a remote site that can be activated in less than twenty-four hours if needed.

Best Practices—Integrity of Data:

- Metadata must be collected about the record and maintained with the record, either embedded in it, or stored separately. Descriptive metadata is used for the indexing, discovery, and identification of a digital resource. At a minimum, descriptive metadata should include creator, date, collector, and description. Land and property transactions should include the grantor/grantee names, title-file, date-file time, book and page, and description. Administrative metadata is information that is needed for the management of the digital object, which includes information regarding ownership, transfer information, access and display, and rights management. Preservation metadata that need to be collected includes the file format, record type, e.g. tax map or correspondence, the operating system, software configurations, the rights/security, and versioning information. For more information see http://www.ncecho.org/ and consult the metadata initiatives section.
- Security measures—Digital Fingerprinting
    - Information can be lost during transmission, migration, or when media breaks down or is corrupted. To ensure that the data does not and has not changed, you should perform a digital fingerprint procedure [e.g. digital certificates, Cyclical Redundancy Checksums or CRCs, and cryptographic hashing algorithms such as a Secure Hashing Algorithm (SHA)]. However, keep in mind that a CRC verifies the transmission of the document but not the document itself. A SHA verifies both the transmission and the information in the document itself. A digital fingerprint is unique to each document and verifies the integrity (unaltered state) of the document. When auditing the information or storage media, reproducing the digital fingerprint can determine if data has been lost. If you employ digital fingerprinting, retain the method by which it was applied so it can be recreated and compared to the original fingerprint.
    - Integrity of the record: If you elect to employ/allow digital fingerprints, you should have a migration strategy in place and a method to verify the fingerprint in the future so that it is preservable and upwardly migratable. As part of your migration strategy, a digital fingerprint should be created at the beginning and at the end of the migration to ensure that the numbers produced from the algorithm are the same. If the two “fingerprints” match, then no error occurred during the transmission or migration.
- Security measures—Authority Rights. If special authority is needed to access the information, indicate who has that authority and the data type (e.g. document or photograph).
- For admissibility of records, the content, context, and structure should be preserved.

Best Practices—System Parameters:

- Document the system that produced the record including the system hardware and software versions used to create the record. Policies and procedures for all aspects of system operation and maintenance, including procurement, data entry, quality control, indexing, corrections, expungement, redaction, back-ups, security, and migration, are all security mechanisms that serve as safeguards to protect against tampering and unauthorized access and printing.
- The following items should be maintained for archival entries:

1. All system equipment specifications.
2. Contact information for manufacturers and vendors.
3. A description of all hardware and software upgrades to the system, including date of maintenance and version of software along with setting change, date, time, and name of operation.
4. Technical and user operation manuals.
5. All policies and procedures related to access to and security of the records.

- Any changes made to the system or the process should be documented.
- The system should be capable of providing audit trails and system security. Effective audit trails can automatically detect who had access to the system, whether staff followed existing procedures, or whether fraud or unauthorized acts occurred or are suspected.
- A migration strategy should be established and implemented for regular recopying, reformatting, and other necessary maintenance to ensure the retention and usability of electronic records throughout their authorized life cycle. Migration needs to maintain the content of the records and any associated metadata required to interpret the records including record format or layout, contextual elements, and the data's relationship to other data.
- Document the controls that monitor the accuracy and authenticity of data, the reliability of hardware and software, and the integrity and security of the system.
- Use open-source software.
- Use preferred file formats for text documents such as rich text format (rtf), .pdf/a, PDF.
- Use preferred file formats for photographic and image documents such as TIFF Group 4 lossless compression, JPEG2000, SQL database.
- Copy immediately onto new media any permanent or archival electronic records stored on media with 10 or more permanent errors per volume.
- Copy all permanent or archival electronic records onto new media before the media is 5 years old. While manufacturer specifications might promise a longer lifetime of media, independent tests show media degradation as early as five years. Additionally, new software technologies usually come to market within five years. Without the software to read the data, it becomes unreadable.

Best Practices—Media Preservation and Storage:

- Select appropriate storage media and environment.
- Store media in environmentally controlled conditions. Humidity should not exceed 50% and should not fall below 30%. Room temperature should be stable at 65 to 75 degrees Fahrenheit. Adhere to the media manufacturer's recommendations for specific environmental conditions in which the media should be stored.
- Never operate drive systems in environments with high levels of airborne particles.
- If using optical media, periodically clean optical media to remove dust and other particulates. In addition, periodically clean drivers that read the media to ensure their operation.
- To protect disks from warping, they should not be subject to pressure and should be stored in an upright position when not in the disk drive.
- For magnetic computer media tapes that contain permanent or archival electronic records, tapes should be rewound under constant tension at least every 2 years. Annually test a 3 percent statistical sample of all volumes, or 10 volumes of each type of

magnetic media, whichever is larger, to identify any loss of data and to discover and correct the causes of data loss;

- Labels for media should include the following identifiers:
    - Creator, date created, division or agency where created, name of agency, unit, and division that is responsible for the records on the disk; hardware, operating system, and software required to access the index or information on the disk; encoding standard and version; model of security or restricted access; sequential number or other specific identifier that identifies the disk in the series of disks used by the system; identification of the disk as master or back-up storage copy; retention dates of the information on the media; data classification as to whether stored off-site, confidentiality of the data, who can access it, who can read the data, and are there different models of confidentiality (e.g. are parts of the record public records while parts of it are confidential?)
- If the disk or other format is too small to include all of the information on the label, then establish a coding system that can be linked back to an index that holds all of the vital information. Documentation relating to the coding system and index must be maintained for as long as it relates to any labeled storage medium.

Electronic document images should be true copies of the documents from which they were made. A true copy is defined as being one that accurately reproduces an original document.

Best Practices—Eye to the Future:

- Practitioners of a trusted digital repository should take measures to keep abreast of and adapt to changing industry standards and technologies to ensure the survivability of the system.

Best Practices—Legal Admissibility Standards:

- *The Uniform Photographic Copies of Business and Public Records as Evidence Act, (UPA) [US 1128-0020-00],* permits the substitution of photographic copies for original documents for judicial or administrative purposes, provided that the copies are produced in the regular course of business and that no laws or regulations require retention of the original documents. Where these conditions are satisfied, the Uniform Photographic Copies of Business and Public Records as Evidence Act permits, but does not mandate, the destruction of original documents. In the case of North Carolina, however, specific exemptions are made, as follows:

- *G.S. § 8-45.1. Photographic reproductions admissible; destruction of originals.*
  (a) If any business, institution, member of a profession or calling, or any department or agency of government, in the regular course of business or activity has kept or recorded any memorandum, writing, entry, print, representation, X-ray or combination thereof, of any act, transaction, occurrence or event, and in the regular course of business has caused any or all of the same to be recorded, copied, or reproduced by any photographic, photostatic, microfilm, microcard, miniature photographic, or other process which accurately reproduces or forms a durable medium for so reproducing the original, the original may be destroyed in the regular course of business unless held in a custodial or fiduciary capacity or unless its preservation is required by law. Such

reproduction, when satisfactorily identified, is as admissible in evidence as the original itself in any judicial or administrative proceeding whether the original is in existence or not and an enlargement or facsimile of such reproduction is likewise admissible in evidence if the original reproduction is in existence and available for inspection under direction of court. The introduction of a reproduced record, enlargement or facsimile, does not preclude admission of the original.

(b) The provisions of subsection (a) of this section shall apply to records stored on any form of permanent, computer-readable media, such as a CD-ROM, if the medium is not subject to erasure or alteration. Non-erasable, computer readable storage media shall not be used for preservation duplicates, as defined in G.S. 132-8.2, or for the preservation of permanently valuable records as provided in G.S. 121-5(d), except to the extent expressly approved by the Department of Cultural Resources pursuant to standards and conditions established by the Department. (1951, ch. 262, s. 1; 1977, ch. 569; 1999-131, s. 1; 1999-456, s. 47(a).)

- *G.S. § 153A-436. Photographic reproduction of county records.*

  (a) A county may provide for the reproduction, by photocopy, photograph, microphotograph, or any other method of reproduction that gives legible and permanent copies, of instruments, documents, and other papers filed with the register of deeds and of any other county records. The county shall keep each reproduction of an instrument, document, paper, or other record in a fire-resistant file, vault, or similar container. If a duplicate reproduction is made to provide a security copy, the county shall keep the duplicate in a fire-resistant file, vault, or similar container separate from that housing the principal reproduction.

  If a county has provided for reproducing records, any custodian of public records of the county may cause to be reproduced any of the records under, or coming under, his custody.

  (b) If a county has provided for reproducing some or all county records, the custodian of any instrument, document, paper, or other record may permit it to be removed from its regular repository for up to 24 hours in order to be reproduced. An instrument, document, paper, or other record may be removed from the county in order to be reproduced. The board of commissioners may permit an instrument, document, paper, or other record to be removed for longer than 24 hours if a longer period is necessary to complete the process of reproduction.

  (c) The original of any instrument, document, or other paper received by the register of deeds and reproduced pursuant to this Article shall be filed, maintained, and disposed of in accordance with G.S. 161-17 and G.S. 121-5. The original of any other county record that is reproduced pursuant to this Article may be kept by the county or disposed of pursuant to G.S. 121-5.

  (d) If an instrument, document, or other paper received by the register of deeds is reproduced pursuant to this Article, the recording of the reproduction is a sufficient recording for all purposes.

  (e) A reproduction, made pursuant to this Article, of an instrument, document, paper, or other record is as admissible in evidence in any judicial or administrative proceeding as the original itself, whether the original is extant or not. An enlargement or other facsimile of the reproduction is also admissible in evidence if the original reproduction is extant and available for inspection under the direction of the court or administrative agency.

  (f) The provisions of this section shall apply to records stored on any form of permanent, computer-readable media, such as a CD-ROM, if the medium is not subject to erasure or alteration. Non-erasable, computer-readable storage media shall not be

used for preservation duplicates, as defined in G.S. 132-8.2, or for the preservation of permanently valuable records as provided in G.S. 121-5(d), except to the extent expressly approved by the Department of Cultural Resources pursuant to standards and conditions established by the Department. (1945, c. 286, ss. 1-7; c. 944; 1951, c. 19, ss. 1-6; 1953, c. 675, ss. 23, 24; 1957, c. 330, s. 3; 1973, c. 822, s. 1; 1999-131, s. 4; 1999-456, s. 47(d).)

- *Rule 1003 of the Uniform Rules of Evidence and Federal Rules of Evidence* provides for admission of duplicate records in evidence unless serious questions are raised about the authenticity of original records from which the copies were made or, in specific circumstances, admitting a copy in lieu of an original is judged unfair. Rule 1003 does not require that duplicate records be produced in the regular course of business. It does not authorize or prohibit destruction of original records.

(This document was modified from an original submitted to the Electronic Recording Council of the Office of the Secretary of State and was adopted by the Council on June 20, 2006 as part of the *North Carolina Electronic Recording Standards*, approved by the Secretary of State on April 18, 2007.)

Sources

CENDI, *Formats for Digital Preservation: A Review of Alternatives and Issues,* http://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf, March 2007.

Center for International Earth Science Information Network (CIESIN), *Guide to Managing Geospatial Electronic Records*. Columbia University, 2005.

Natoli, James G., New York State Office for Technology. *"Governor's Task Force on Information Resource Management, Technology Policy 96-10"* 1996.

Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1 Blue Book, January 2002.

Rothenberg, Jeffrey, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Council on Library and Information Resources: Commission on Preservation and Access Digital Libraries, 1998.

North Carolina Exploring Cultural Resources, *NC ECHO Dublin Core Implementation Guidelines,* Raleigh, North Carolina, April 24, 2004.

State Archives Department, Minnesota Historical Society, *XML for Information Management,* St. Paul, Minnesota, October 2002.

ADDENDUM A

Glossary of Terms

- **Authentication**: The act of tying an action or result to the person claiming to have performed the action. Authentication generally requires a password or encryption key to perform, and the process will "fail" if the password or key is incorrect.

- **Digital signature**: A complex string of electronic data that is embedded in an electronic document for the purposes of verifying document integrity and signer identity. A mainstay of the Public Key Infrastructure (PKI), digital signatures are the most effective method for ensuring non-repudiation for digital documents.

- **Digitized signature**: A representation, e.g. a scanned version, of a person's handwritten signature, existing as a computerized image file. Digitized signatures are just one of several types of electronic signatures, and have no relation to digital signatures.

- **Document type definition** (DTD): A document created using the Standard Generalized Markup Language (SGML) that defines a unique markup language (such as XHTML or XML). A DTD includes a list of tags, attributes, and rules of usage.

- **Electronic signature**: Any of several methods that links a person to a document or action using electronic data. According to electronic signature laws in the U.S. (including the federal Electronic Signatures in Global and National Commerce Act, E-SIGN, and the Uniform Electronic Transactions Act, UETA), any embedded electronic element can serve as a signature if a person embeds it with the intent to sign.

- **Encrypt**: To apply an encryption key to a message in order to make it unreadable in an effort to prevent unintended use of the information.

- **Extensible Markup Language** (XML): A computer language used to create markup languages. XML allows developers to specify a document type definition (DTD) or schema in order to devise new markup languages for general or specific uses.

- **Hash function**: A mathematical algorithm that takes an electronic document and creates a document fingerprint. The document fingerprint is much smaller than the original document, and does not allow the reconstitution of the original document from the fingerprint. A slightly different document, processed through the same hash function, would produce a very different document fingerprint. A hash function helps to secure data by providing a way to ensure that data are not compromised.

- **Metadata: "**Metadata is commonly defined as 'data about data.' Metadata is frequently used to locate or manage information resources by abstracting or classifying those resources or by capturing information not inherent in the resource. Typically metadata is organized into distinct categories and relies on conventions to establish the values for each category. For example, administrative metadata may include the date and source of acquisition, disposal date, and disposal method. Descriptive metadata may include information about the content and form of the materials. Preservation metadata may record activities to protect or extend the life of the resource, such as reformatting. Structural metadata may indicate the interrelationships between discrete information

resources, such as page numbers." (Source: Richard Pearce-Moses: *A Glossary of Archival & Records Terminology* Society of American Archivists, 2005*)*

- **Proprietary**: Indicates that software or other employed technology is owned or controlled exclusively by the vendor. These solutions are not transferable to other systems and must be used only on the vendor's systems.

- **Signature authentication**: The process by which a digital signature is used to confirm a signer's identity and a document's validity.

- **Signed digital document**: An electronic document that includes an embedded digital signature. The digital signature contains an encrypted document fingerprint that allows anyone receiving the document to verify its validity using the process of signature authentication.

- **Tagged information file format** (TIFF): An image file format commonly used for photos, scanned documents, or other graphics. TIFF images are graphics that are made up of individual dots or pixels. Files in the TIFF format are distinguished by a .tif filename extension.

- **XML**: Short for **Extensible Markup Language** (**XML**) is a general-purpose markup language. It is classified as an extensible language because it allows its users to define their own tags. Its primary purpose is to facilitate the sharing of structured data across different information systems, particularly via the Internet. It is used both to encode documents and serialize data. XML can be used to store any kind of structured information, and to enclose or encapsulate it in order to pass the information between different computing systems which would otherwise be unable to communicate.