Issued by the EPA Chief Information Officer,
Pursuant to Delegation 1-19, dated 07/07/2005

Digitization (Scanning) Standard

1. PURPOSE

To establish a standard for capturing digitized (scanned) content from paper, microfilm and/or microfiche from Agency documents and records in Agency content repositories or other designated digital storage environments. The standard is designed to enhance the efficiency of Agency digitization efforts and ensure that the quality of digitized documents meets intended uses.

2. SCOPE

The standard covers digitization efforts across the Agency and applies to all EPA programs, regions, laboratories and offices. The standard shall be used by owners of existing systems and applications that are currently digitizing documents within the scope of their operating authority (e.g., the Superfund Enterprise Management System, the Federal Docket Management System, the Correspondence Management System, etc.). The standard is intended to supplement other EPA information management policies, procedures and standards that focus primarily on operations for digitizing documents and records for delivery to Agency document/records management applications. The standard may also be relevant to and considered when initially capturing and managing electronic information.

3. AUDIENCE

The audience for the standard includes all EPA organizations, officials and employees, as well as contractors, grantees and other agents of EPA that digitize Agency-owned paper- or microform-based records and documents.

4. BACKGROUND

Several forces within the federal government are uniting to spur digitization. Drivers for digitization include the increased need for transparency and accessibility to information, the desire for enhanced mobility, and the desire to reduce the physical footprint of government office space. Other drivers for digitization include the National Archives and Records Administration (NARA) and the Office of Management and Budget (OMB) Memorandum M-19-21, with the following goals:

- Goal 1.1, Requiring that all permanent electronic records be managed electronically by December 31, 2019, to the fullest extent possible for eventual transfer and accessioning to NARA in an electronic format.
- Goal 1.2, Requiring all permanent records in federal agencies be managed in an electronic format with appropriate metadata by December 31, 2022.

- Goal 1.3, Requiring all temporary records in federal agencies to be managed in electronic format or stored in commercial records storage facilities by December 31, 2022.
- Goal 2.4, Requiring NARA to no longer accept transfers of permanent or temporary records in analog formats (hardcopy, microfilm and microfiche) and only accept records in electronic format and with appropriate metadata, after December 31, 2022.

Benefits from the standard include:

- Productivity improvement due to enhanced access to Agency records/documents;
- Reduction in the time and effort required to search for documents and records needed for a variety of regulatory and mission-related reasons;
- Decrease in the number of filing errors and the volume of duplicate content;
- Reduction in and better management of the overall volume of hard-copy (paper) information;
- Easier data sharing among information systems across the enterprise; and
- Enhanced identification, sharing and use of Agency information resources by EPA's information customers and stakeholders.

The electronic management of digitized documents and records will also result in subsequent reductions in the costs associated with paper-based documents and records. The standard is thus designed to:

- Support the migration from hard-copy/paper-based documents to electronic documents;
- Integrate and standardize the digitization process as part of the records life cycle;
- Leverage existing Agency investments in the EPA Enterprise Architecture (e.g., Documentum® and its enterprise storage environment, scanners, etc.), Enterprise Content Management (ECM) systems such as the Correspondence Management System (CMS), Federal Docket Management System (FDMS) and Superfund Enterprise Management System (SEMS), and Enterprise Information Management (EIM);
- Serve as a framework into which additional program-specific standards and workflows can be incorporated, based upon the needs of the business units; and
- Establish the basic standard business practices necessary to satisfy the requirements of the Federal Rules of Evidence, the Federal Records Act, and other authorities, policies and procedures under which the Agency must operate, such as NARA and known best practices

5. AUTHORITY

- Clinger-Cohen Act (also known as Information Technology Management Reform Act of 1996) (Pub. L. 104-106, Division E)
- Paperwork Reduction Act of 1980, as amended by the Paperwork Reduction Act of 1995 (44 U.S.C. Chapter 35)
- Government Paperwork Elimination Act of 1998 (Pub. L. 105-277, Title XVII)

- United States vs. Russo, 480 F.2d 1228, 1239 (6th Cir. 1973)
- Presidential Memorandum: Managing Government Records, November 28, 2011
- Presidential Memorandum: Building a 21st Century Digital Government, May 23, 2012
- Executive Order – Making Open and Machine Readable the New Default for Government Information, May 9, 2013
- NARA/OMB Directive M-12-18: Managing Government Records, August 24, 2012 (Superseded by M-19-21)
- OMB Circular No. A-130: Management of Federal Information Resources
- OMB Memorandum M 10-06: Open Government Directive, December 8, 2009
- OMB Memorandum M-13-13: Open Data Policy - Managing Information as an Asset May 9, 2013
- CIO 2130.1: Section 508: Accessible Electronic and Information Technology February 20, 2014 (http://intranet.epa.gov/oei/imitpolicy/qic/ciopolicy/2130.1.pdf)
- NARA/OMB Memorandum, M-19-21: Transition to Electronic Records, June 8, 2019

6. STANDARD

EPA programs, regions, laboratories and offices are directed to:

- Use the digitization standard for capture of hard-copy documents and records in Agency content repositories or other designated storage environments (e.g., CMS, FDMS), where use does not jeopardize existing standard business practices; and
- Incorporate the digitization standard into documented standard operating procedures (SOPs) to ensure consistency across the Agency and establish the framework for legally-defensible standard business practices for digitization. For additional information on digitization SOPs, please refer to the related EPA Information Directive – Digitization (Scanning) Procedure
- Ensure the parameters inform the equipment selection, as well as the decision to perform the work at EPA or through a contract vehicle

Hardware ("brand neutral") Standard

A. Low volume scanner standard

The standard designates the acceptable scanner device for low volume (i.e., incidental/infrequent use for small-batch jobs < 25 pages) applicable for the scanning of standard office paper materials only:

- Desktop/stand-alone flatbed scanners;
- Multi-function copier/printer machines;
- All-in-one scanners/printers; and
- Wide-format scanners for oversized documents, up to 34 in. x 44 in. (i.e., page measurement standards ISO-A0 and ANSI-E).

B. High volume scanner standard

The standard designates the acceptable scanner devices for high volume (i.e., frequent

use for large-batch jobs >25 pages) applicable for the scanning of standard office paper materials only:

- 1,000 page/hour minimum throughput;
- Compatible with Enterprise Capture Software standard (see software standard below);
- ISS- and Twain-driver compatible;
- Native (on board), or compatible with, Kofax Virtual Re-Scan® (VRS®) quality-enhancing production software, or alternatively, the Captiva®/Input Accel® Image Quality Checks feature;
- Sheet size capability from 2.05 in. x 2.91 in. (i.e., page measurement standard ISO-A8) up to 11 in. x 17 in. (i.e., page measurement standards ISO-A3 and ANSI-B);
- Duplex (2-side scanning) capability; and
- Color, gray-scale and monochrome capability.

C. Film digitizers standard (e.g., microform, microfilm, slides, etc.)
The standard directs users to address the following characteristics that may influence the digitization approach or affect the digital image quality:

- The type and volume of the materials to be digitized;
- Text quality and clarity on the microfilm;
- The quality of the original capture of the film (lack of focus, uneven lighting, page curvature, gutter shadows, etc.);
- Variations in density between exposures;
- The reduction ratio of the film;
- Resolution and the ability to detect detail on the film; and
- The condition of the film itself (scratches, etc.).

Digitizing and capture software standard

The standard here applies only to new acquisitions or upgrades to the software already in use in the Agency. They are not intended to require wholesale replacement of software used now or in the past.

D. Low volume digitizing and software applications standard

- Stand-alone (non-networked) usage:
  - Manufacturer-supplied capture software;
  - Manual submission of output to Enterprise Capture Software (see below); and
  - Network-attached usage:
    - Integrated with Enterprise Capture Software (see below).

E. High volume digitizing and software standard

- Enterprise Capture (high volume, as defined in the high volume scanner standard above) network-attached usage:
  - EMC Captiva® (InputAccel®)

- Kofax Capture (Multiple server-level installations across the Agency)

F. Analog or film based digitizing and software standard (e.g., microform, microfilm, slides, etc.)
- Manufacturer-supplied capture software

Content digitized file format standard

G. PDF/A-1 file format standard (Portable Document Format/Archive)
- Preferred format for documents that are primarily textual in nature;
- Image-over-text content indexing (a.k.a., optical character recognition, or OCR);
- Optimized for Internet/Web streaming;
- NARA preferred specification for transfer to Archive:
  - ISO 19005-1:2005 electronic document file format for long-term preservation – part 1: Use of PDF 1.4 (PDF/A-1): (https://www.iso.org/standard/38920.html)
  - Not the preferred output for non-networked scanning of textual documents where that output should be passed on to Enterprise Capture software for processing (see the TIFF file format standard below)
  - Not the preferred output for non-textual materials such as graphics, maps and photographs (see the JPEG file format standard below)

H. TIFF file format standard (formerly Tagged Image File Format)
- Preferred format for low volume, stand-alone document scanning where the TIFF file can be passed on (manually or via automated workflow) to Enterprise Capture software for additional processing such as OCR, image enhancement, conversion to PDF/A, etc.NARA specification for transfer to Archive:
  - TIFF Revision 6.0 Final – June 3, 1992 Adobe Systems, Inc. (https://www.adobe.io/content/dam/udp/en/open/standards/tiff/TIFF6.pdf)

I. JPG file format standard (Joint Photographic Experts Group)
- Preferred format for non-textual documents that are primarily graphical (image) in nature, e.g., maps, photos;
- Compression should not result in an image quality of 10% or less than the original image to preserve image quality while minimizing file size;
- NARA specification for transfer to Archive:
  - ISO/IEC 15444-1:2004 Information technology – JPEG 2000 image coding system: Core coding system (https://www.iso.org/standard/37674.html)

Content image standard

J. Image resolution standard
- Predominately textual documents:
  - Good-to average quality originals – Bi-tonal (2-bit), scanned at a minimum of 300 pixels per inch (ppi), up to 600 ppi
  - Average-to-poor quality originals – Low inherent contrast, staining or fading,

e.g., carbon copies, thermofax, NCR/carbonless paper or documents with handwritten annotations or other markings – Bi-tonal (2-bit), scanned at a minimum of 400 ppi

- Predominately textual documents of good-to-poor quality with gray-scale or color illustrations, photos or text containing color important to interpretation or content – 24-bit RGB (Red, Green, Blue), scanned at 300-400 ppi
- Non-textual (or minimal text content) graphics, illustrations, photos, charts and maps – 24-bit RGB (Red, Green, Blue), scanned at 300-400 ppi

NOTE: Depending upon the type of scanner and capture software used, it may be useful and more convenient to simply apply the settings for 24-bit RGB (Red, Green, Blue), scanned at 300-400 ppm (as described above) as a default for all document scanning.

K. Skew standard

- Three degrees (3^0) or less
- When using scanners so equipped, the skew standard setting should be applied to the Kofax Virtual Re-Scan® (VRS®) quality-enhancing production software, or alternatively, the Captiva®/Input Accel® Image Quality Checks feature, in order to optimize batch processing and to ensure the skew standard is monitored by the software

L. Speckle standard

- Five percent (5%) or less
- When using scanners so equipped, the speckle standard should be applied to the Kofax Virtual Re-Scan® (VRS®) quality-enhancing production software, or alternatively, the Captiva®/Input Accel® Image Quality Checks feature, in order to optimize batch processing and to ensure the speckle standard is monitored by the software

M. Contrast and brightness standard

- Due to variances in scanners and software, each digitization installation should run test batches of documents to be digitized to determine the capture software contrast and brightness setting calibrations that are needed for optimum document viewing, utility, and production software functionality
- When using scanners so equipped, the settings determined from the operations described in the above bullet should be applied to the Kofax Virtual Re-Scan® (VRS®) quality-enhancing production software, or alternatively, the Captiva®/Input Accel® Image Quality Checks feature, in order to optimize batch processing and to ensure the minimum contrast and brightness parameters are monitored by the software.

Output information standard

N. Content indexing standard (a.k.a., Optical/Intelligent Character Recognition – OCR, ICR)

- Only with human review and re-keying can 100% content indexing accuracy for scanned documents be achieved. For truly effective, efficient and accurate retrieval of digitized content from content management systems, content indexing must be

supplemented by cataloguing (indexing) documents for metadata-based searches, as described in the cataloguing and categorization standard below.

- All textual documents should be content indexed during the digitization/capture process
- Whenever possible, content indexing should be accomplished using the Enterprise Capture software standard described above. For low volume scanners, this may require passing TIFF file output to the Enterprise Capture software, utilizing the Agency's data network(s), secure Web portal, or via secure email

O. Cataloguing and categorization standard (metadata indexing)

- Associating metadata with an imaged (scanned) file is necessary to meet the NARA's definition of a high-quality "production master image." Additionally, 100% accuracy in content indexing (see content indexing standard above) is rarely achieved during scanning operations. This necessitates the cataloguing of scanned content in order to maximize the power, effectiveness and accuracy of enterprise information search/retrieval tools
- Digitized documents should generally be catalogued using the Agency's Information Standard: Enterprise Information Management (EIM) Minimum Metadata Standard (see Section 8 below) – or depending upon the source and type of document, using other appropriate Agency metadata standards – and more granular document taxonomies, as registered in the Agency's data resource registries and repositories

Quality standard

P. Quality assurance and quality control

Quality control during the digitization process, and quality assurance of digitized content, is critical to ensuring the integrity, reliability and utility of the content for uses that support the Agency's mission.

- Some basic QA and QC operations should be incorporated in the capture process through the use of quality-enhancing production software tools such as VRS (Kofax) and Captiva's Image Quality Checks feature (see the output information standards above)
- To ensure an effective and consistent approach to QA and QC, digitization/capture should conform to a formal Agency-level Quality Assurance Plan (QAP), developed and established for Agency digitization operations, pursuant to the CIO 2106: Quality Policy; Procedure for Quality Policy and NARA's Digitization Standards found at 36 CFR Chapter XII, Subchapter B, Part 1236, Subpart D.

NOTE: For waivers to the Content Parameters, see Section 10.

7. ROLES AND RESPONSIBILITIES

The roles and responsibilities with respect to the digitization standards include:

The Chief Information Officer (CIO)

- Lead Agency-wide implementation of the Digitization Standard as part of the

overall framework of CIO Policies

- Facilitate the process for appropriate business organizations to incorporate the standards into their organization and operations
- Manage the Senior Advisory Council process to update the standards and associated policies and procedures, and propose new information policies, procedures and standards as needed
- Authorize formal information calls for updates or reviews of the standard, as appropriate
- Grant waivers to selected provisions of the standard for sufficient cause, or delegate waiver authority

Senior Advisory Council (SAC)

- Advise and assist the Chief Information Officer in developing and implementing the Agency's quality and information goals and policies
- Review updates to the Digitization Standard and associated policies and procedures, and propose new information policies and procedures as needed
- Review any progress reports provided and address successes, as well as Agency-wide challenges, for the effective implementation of the standard
- Endorse enterprise-wide information investments, coordinating with Agency Investment Oversight Boards, as appropriate

Senior Information Officials (SIOs)

- Implement the standard within their organizations
- Apprise the SAC of major digitization issues within their offices
- Ensure that the information technology used and managed by their organizations supports their business needs and mission and helps to achieve strategic goals
- Ensure Enterprise Architecture compliance of solution architectures
- Review, concur, and advise on waivers to the standards, typically through participation on the Information Technology Operations Planning Committee (IOPC)

Records Liaison Officers (RLOs)

- Participate in the development and maintenance of digitization standard operating procedures, as appropriate, for relevant programs, regional offices, laboratories, etc.
- Support and implement the Digitization Standard, and related technical specifications and standard operating procedures
- Work with records, document and content owners/generators to plan and manage the life cycle of the digitized materials
- Oversee the implementation of such plans throughout the life cycle of the digitized materials
- Coordinate with Information Management Official(s) and provide outreach, support, and technical assistance as appropriate to ensure the proper implementation of the standard

Information Management Officials (IMOs)

- Review, concur on or approve acquisition of digitization technologies to ensure compliance with the standard
- Review Agency digitization policy documents, as appropriate. Proposals to set new standards and procedures should be submitted to the appropriate group under the IT/IM Governance model, following the most current CIO Policy Review procedures
- Ensure that staff and contractors are aware of the standard, and related technical specifications and standard operating procedures
- Ensure that employees, senior environmental enrollees, and contractors are aware of their responsibilities regarding digitization
- Review and/or certify compliance with the standard and other Agency digitization policies and procedures, as appropriate

All EPA employees and agents

- Use the standard to manage Agency-owned unstructured information in content repositories

8. RELATED INFORMATION

- CIO 2105.0: Policy and Program Requirements for the Mandatory Agency-Wide Quality System, May 5, 2000
- CIO 2106.0: Quality Policy, October 20, 2008
- CIO 2106-P-01.0 Procedure for Quality Policy, October 20, 2008
- CIO 2122.1: Enterprise Architecture Policy December 21, 2017
- CIO 2122-P-03.0: Information Technology Infrastructure Standard Procedure, October 1, 2010
- CIO 2133.0: Data Standards, June 28, 2007
- CIO 2155.4: Interim Records Management Policy, August 22, 2018
- CIO 2135-P-01.0: Enterprise Information Management Policy Cataloguing Information, March 3, 2015
- CIO 2135-S-01.0: Enterprise Information Management (EIM) Minimum Metadata Standards, March 3, 2015
- EPA Geospatial Metadata Technical Specification, August 2016
- CIO 2131.0: National Geospatial Data Policy, August 24, 2005

9. DEFINITIONS

Content: The intellectual substance of a document, including text, data, symbols, numerals, images and sound. (Society of American Archivists)

Content Management: The capability to manage and track the location of, and relationships among, content within a repository (AIIM International)

Content Repository: A database that securely stores electronic content and associated metadata with management controls

Data Resource Registry: An application which stores metadata for querying, and which can be used by any other application in the network with sufficient access privileges. A registry is an index of a data or metadata repository which is made up of all the data providers' data and reference metadata sets within a community, distributed across the Internet or similar network. The registry services are not concerned with the storage of data but rather with providing visibility of the data and reference metadata, and information needed to access the data and reference metadata. (http://stats.oecd.org/glossary/detail.asp?ID=7078)

Data Resource Repository: "A central place where data are stored and maintained. It can be a place where multiple [data,] databases or files are located for distribution over a network, or a repository can be a location that is directly accessible to the user without having to travel across a network." (http://www.webopedia.com/TERM/R/repository.html)

Document: Information set down in any physical form or characteristic. A document may or may not meet the definition of a record. (DOD 5015.2-STD)

Enterprise: An organization (or cross-organizational entity) supporting a defined business scope and mission. An enterprise includes interdependent resources (e.g., people, organizations, and information technology) that must coordinate their functions and share information in support of a common mission (or set of related missions)

Enterprise Content Management: The strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization's unstructured information, wherever that information exists. (AIIM International, http://www.aiim.org/What-is-ECM-Enterprise-Content-Management)

Guidance: A non-mandatory compilation of advice, examples, best practices or past experience. Guidance supplements procedures. (EPA Web Governance and Management Policy)

Information: For purposes of the standards, information means any communication or representation of knowledge such as facts or content, in any medium or form, including, but not limited to, textual, numerical, graphic, cartographic, narrative, or audiovisual forms. (OMB Information Quality Guidelines)

Metadata: Data describing stored data; that is, data describing the structure, data elements, interrelationships, and other characteristics of electronic records. (DOD 5015.2)

Organization: A company, corporation, firm, enterprise, or institution, or part thereof, whether incorporated or not, public or private, that has its own functions and administration. In the context of the standards an EPA organization is an office, region, national center, or laboratory

Policy: A high-level statement about an Agency requirement designed to influence and determine decisions, actions, and other matters. It is usually driven by statute, Executive Order, the mandate of an oversight agency or Congress, or the head of the organization. (EPA Web Governance and Management Policy)

Procedure: The required steps, courses of action, or processes needed to accomplish or satisfy a policy. It provides a basis for assuring consistent and acceptable minimum levels of quality, performance, safety and reliability. Standards usually are included in, or accompany, procedures. (EPA Web Governance and Management Policy)

Senior Advisory Council (SAC): The SAC consists of high-level managers from each Region and program office, typically at the Deputy Assistant Administrator level. The SAC's primary focus is on addressing and resolving intra-Agency cross-media, cross-program, and interdisciplinary information technology/information management and related policy issues

Quality Assurance: A management or oversight function that deals with setting policy and running an administrative system of management controls that cover planning, implementation, review, and maintenance to ensure products and services are meeting their intended use

Quality Control: The overall system of technical activities that measure the attributes and performance of a process, item, or service against defined standards to verify that they meet the stated requirements established by the customer; operational techniques and activities that are used to fulfill requirements for quality.

Record(s): All recorded information, regardless of physical form or characteristics, made or received by an agency of the United States government under federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the government or because of the informational value of data in them. (44 U.S.C. §3301)

Records Management: The planning, controlling, directing, organizing, training, promoting and other managerial activities involved with respect to records creation, records maintenance and use, and records disposition in order to achieve adequate and proper documentation of the policies and transactions of the federal government and effective and economical management of agency operations. (36 CFR §1220.14)

Standard: Universally or widely accepted, agreed upon, or established means of determining what something should be. Major classifications of this term include: (1) Material or substance whose properties are known with a level of accuracy that is sufficient to allow its use as a physical reference in calibrating or measuring the same properties of another material or substance. (2) Concept, norm, or principle established by agreement, authority, or custom, and used generally as an example or model to compare or measure the quality or performance of a practice or procedure. (3) Written definition, limit, or rule approved and monitored for compliance by an authoritative agency (or professional or recognized body) as a minimum acceptable benchmark. This is the usual meaning of the plural term (standards) (http://www.businessdictionary.com/definition/standard.html)

Taxonomy: A hierarchical structure of information components, any part of which can be used to classify a content item in relation to other items in the structure (from “The Challenges of Building Enterprise Content Taxonomies and the Role of Classification Technologies in Maintaining their Effectiveness,” Reginald J. Twigg, PhD, IBM

Corporation, 2007)

Web: A system of Internet servers that support specially formatted documents. The documents are formatted in a markup language called HTML (Hypertext Markup Language) that supports links to other documents, as well as graphics, audio, and video files

10. WAIVERS

Consistent digitization standards are critical to facilitating the exchange, use and integrity of the Agency's unstructured information. For this reason, waivers to the standards are rare and will be considered on a case by case basis.

Waiver Process: The Agency's CIO may grant waivers to selected provisions of the standards for sufficient cause. The CIO may re-delegate the authority (for example, to the Electronic Content Subcommittee of the Senior Advisory Council).

Requests: Requests for waivers to specific provisions of the standards must conform to the appropriate OMS-EI waiver procedures, and must contain 1) identification of the standards provision; 2) a listing of reasons why the standards cannot be applied or maintained; 3) an assessment of impacts resulting from non-compliance; and 4) a memorandum to the CIO originating at the Office Director level (or equivalent) responsible for the information in question, through the SIO or other senior manager.

Notification: The CIO will notify the requesting office in writing of the disposition of the waiver within 60 days of receipt.

11. MATERIAL SUPERSEDED

CIO 2155-S-01.0: Document Digitization (Scanning) Standards, August 4, 2015
NARA/OMB Directive M-12-18: Managing Government Records, August 24, 2012

12. CONTACTS

For questions about this Standard, please contact the U.S. EPA National Records Management Program, (202) 566-1494.

VAUGHN NOGA

Digitally signed by VAUGHN NOGA
Date: 2020.08.18 08:07:18 -04'00'

Vaughn Noga
Chief Information Officer and
Deputy Assistant Administrator for Environmental Information
U.S. Environmental Protection Agency