
CONFERENCE PAPER

Developing Criteria to Establish Trusted Digital Repositories

John Faundeen

U.S. Geological Survey, Earth Resources Observation and Science Center, Sioux Falls, SD 57198, US
faundeen@usgs.gov

This paper details the drivers, methods, and outcomes of the U.S. Geological Survey's quest to establish criteria by which to judge its own digital preservation resources as Trusted Digital Repositories. Drivers included recent U.S. legislation focused on data and asset management conducted by federal agencies spending \$100M USD or more annually on research activities. The methods entailed seeking existing evaluation criteria from national and international organizations such as International Standards Organization (ISO), U.S. Library of Congress, and Data Seal of Approval upon which to model USGS repository evaluations. Certification, complexity, cost, and usability of existing evaluation models were key considerations. The selected evaluation method was derived to allow the repository evaluation process to be transparent, understandable, and defensible; factors that are critical for judging competing, internal units. Implementing the chosen evaluation criteria involved establishing a cross-agency, multi-disciplinary team that interfaced across the organization.

Keywords: Trusted; Digital; Repository; USGS; Preservation; Criteria

Background

As the Nation's largest water, Earth, and biological science and civilian mapping agency, the U.S. Geological Survey (USGS) collects, monitors, analyzes, and provides scientific information about natural resource conditions, issues, and problems. The diversity of the USGS's scientific expertise enables the organization to carry out large-scale, multi-disciplinary investigations and provide impartial scientific information to resource managers, planners, and decision makers.

Drivers

Since the inception of the USGS in 1879, the agency has maintained comprehensive internal and external policies and procedures for ensuring the high quality and integrity of its generated scientific interpretations and products. The documented guidance has led to the USGS' reputation for scientific excellence and objectivity. As new technologies developed and research became more digital, the USGS had to create and adopt new policies. In 1993 the first internal policies were instituted requiring preservation of digital assets. Ten years later the USGS established the web-based USGS Publications Warehouse, its first digital library.

Existing scientific policies and procedures were updated in 2006 and are now known as the USGS Fundamental Science Practices (FSP). These changes established a set of consistent practices, philosophical premises, and operational principles to serve as the foundation for research and monitoring activities related to USGS science.

In January 2009, the Director of the USGS announced the establishment of a Fundamental Science Practices Advisory Committee (FSPAC), which would be responsible for addressing pending and new FSP issues (including previously unresolved issues), fielding questions and concerns about FSP from scientists and managers, and developing recommendations for resolving issues. The FSPAC ensures that the USGS continues to produce high quality, objective science information products by creating guidance on conducting science projects and establishing review processes. In 2012, the FSPAC established a Data Preservation

Sub-Committee (Sub-Committee) to identify and guide USGS science data stewardship, preservation, and documentation requirements. This effort resulted in the 2015 USGS policy entitled “Fundamental Science Practices: Preservation Requirements for Digital Scientific Data”.

These policies reflected the push for open data and access occurring on a national scale. On February 22, 2013, the Office of Science and Technology Policy (OSTP) issued a memorandum, *Increasing Access to the Results of Federally Funded Scientific Research*, which called on all federal agencies with annual research and development (R&D) outlays of more than \$100 million to develop a plan increasing public access to the direct results of federally funded scientific research, including specifically peer-reviewed publications and digital data. In addition, on May 9, 2013, the Office of Management and Budget (OMB) released Memorandum M-13-13, *Open Data Policy-Managing Information as an Asset*. Individually and collectively, these directives established the mandates for the U.S. Federal Government to transform data and information into useable and accessible digital artifacts and promote and accelerate their release (subject to certain limitations imposed by privacy, confidentiality, and national security considerations).

Because 74 percent (\$686 million) of the Department of the Interior's total R&D budget is allocated to the USGS, the bureau was responsible for developing a plan to comply with the OMB memorandum. The USGS Plan focused specifically upon the USGS' public access activities, policies, and plans as they affect both intramural and extramural research and development activities. It also required that data be stored in a USGS Trusted Digital Repository. Because of their aforementioned research on preservation policies, the Sub-Committee was tasked with establishing standards for evaluating USGS data repositories and their trustworthiness.

Methods

Comprised of volunteer staff representing the fields of archival science, digital libraries, computer science, information technology, publishing, and the information sciences, the Sub-Committee evaluated several existing criteria developed by other organizations in order to identify elements relevant to the USGS's pursuit of establishing *Trusted Digital Repositories*. The USGS sought methods that would be considered transparent, authoritative, and scalable allowing minimal to wide interagency use. The criteria and approach also had to be applicable to the USGS in that the anticipated level of effort and the direct costs associated with utilizing a specific implementation could be attainable. Lastly, the USGS sought an approach that offered certifications from a reputable organization. The individual criteria sets examined are outlined in the sections that follow.

U.S. Federal RIM Maturity Model

The first criteria reviewed was the U.S. Federal Records and Information Management Program Maturity Model (JWG FRC and NARA 2014), which included categories such as *Management Support & Organizational Structure, Policy, Standards, & Guidance*, and RIM Program Operations. The categories were then further sub-divided as illustrated below:

- Management Support and Organizational Structure
 - Strategic Planning
 - (a) Management and leadership incorporate RIM as a strategic element of the agency's/component's business and mission.
 - (b) RIM program has strategic goals and objectives.
 - (c) Management places RIM within the part of the agency/component that provides RIM visibility, authority, and sufficient resources.
 - Leadership and Management
 - (a) Leadership and management at all levels endorse the RIM program.
 - (b) Leadership and management consider records and information valuable assets.
 - (c) Agency/component assigns authority and delegates responsibility to personnel with skill sets that align with assigned RIM activities.
 - Resources (Agency/component provides:)
 - (a) Appropriately qualified and trained RIM staff;
 - (b) Sufficient numbers of dedicated staff to meet agency needs for program implementation;
 - (c) Funding for continuing education for RIM staff; and
 - (d) Sufficient funding for services, equipment, technology and acquired resources.
 - Awareness (The extent to which the agency/component:)
 - (a) Has an established method to inform all personnel of their records management responsibilities in law and policy;

- (b) Has developed a communications program that promotes awareness; and
 - (c) Continuously provides up-to-date RIM policy and guidance to all personnel.
- Policy, Standards, and Governance
 - Policy, Standards, and Governance Framework
 - (a) Agency/component assigns responsibility for developing RIM policy, standards, and governance.
 - (b) RIM policy, standards, and governance are documented in an understandable manner.
 - (c) RIM policy, standards, and governance are based upon legislative and statutory regulatory requirements and professional standards.
 - Compliance Monitoring
 - (a) Performance measures and goals are established at the agency/component and program levels.
 - (b) The agency/component has mechanisms in place to monitor and review compliance with RIM policy, standards, and governance.
 - (c) Compliance is measured and reported (internal audits, reviews, and evaluation).
 - Risk Management
 - (a) Agency/component identifies and analyzes internal and external risk to agency/component records and information.
 - (b) Agency/component determines who is best to manage or mitigate the risk and what specific actions should be taken.
 - (c) Agency/component monitors the implementation of actions to management or mitigate risk.
 - Internal Controls are defined as: control activities or processes that provide a reasonable assurance of the effectiveness and efficiency of operations and compliance with RIM policies and practices such as approvals, authorizations, verifications, reconciliations and segregation of duties. The agency/component:
 - (a) Identifies and develops internal controls; and
 - (b) Uses internal controls to ensure compliance with RIM policies, standards, and governance.
 - RIM Program Operations
 - Lifecycle Management
 - (a) Records and information are managed throughout the lifecycle: creation/capture, classification, maintenance, retention, and disposition.
 - (b) Records and information are identified, classified using a taxonomy, inventoried, and scheduled.
 - Retrieval and Accessibility
 - The level to which records and information are easily retrievable and made accessible when needed for agency/component business.
 - Integration
 - (a) RIM is integrated into agency/component-wide business processes.
 - (b) Recordkeeping requirements are integrated into information systems and contracted services.
 - (c) RIM staff participate in system acquisition, development, and/or enhancements.
 - Security and Protection
 - (a) Agency/component has policies in place to protect records and information from internal and external threats.
 - (b) Agency/component has systematic identification and protection of records and information essential for an emergency or Continuity-of-Operations (COOP) event.
 - (c) Agency/component provides guidance on the handling of records and information exempt from disclosure.
 - (d) Agency/component has access controls and safeguards for security classified information as well as other types of restricted information.
 - Training
 - (a) Agency/component ensures that all staff are trained on their records management responsibilities.
 - (b) Agency/component supports professional development for RIM personnel.

This model was found to be extremely comprehensive to the point of even including training elements for staff on records and information management. The model's detail and thoroughness also led the USGS to perceive that implementing such a scheme would be somewhat burdensome. Additionally, the scheme was not intended to provide a certification process.

Digital Curation Centre Checklist for Evaluating Data Repositories

Another criterion examined was compiled by the United Kingdom's Digital Curation Centre entitled, "Where to keep research data: DCC Checklist for Evaluating Data Repositories" (Whyte 2015). This checklist was built around the following questions:

- Is a reputable repository available?
 - Is it recognized in your research domain?
 - Is it recommended by a funder, journal, or learned/professional society?
 - Is it certified as compliant with an appropriate international standard?
- Will the repository take the data you want to deposit?
 - Will it accept any research data, without specifying any categories that exclude what you have to deposit?
 - Does it have a particular focus on data similar to that which you have to deposit, without excluding it on other categories?
 - Does it have an international reputation in your domain, or for publishing data similar to that which you have to deposit, without excluding it on other categories?
- Are the repository's terms and conditions acceptable?
 - By depositing and agreeing to the terms and conditions you will not be contravening relevant institutional policy (e.g. ethics).
 - By depositing and agreeing to the terms and conditions you will not be in breach of copyright, or any contract covering your research, e.g. the grant conditions or a consortium agreement.
 - Anything deposited that is not publicly accessible can be retrieved by the institution in response to an external request that is valid under Freedom of Information legislation.
 - Anything deposited that includes personal data is not stored outside the [United States] Economic Area unless in a legal jurisdiction that ensures an adequate level of protection for the rights and freedom of data subjects in relation to the processing of personal data.
 - It offers a range of licensing options at least one of which is compatible with your institution's policy, and the terms of any contract or agreement with the research funder, partners or participants.
 - It does not require exclusive rights to be transferred.
 - It makes access conditions clear to data users.
 - As a depositor you can identify users to whom file access & edit/delete permission may be restricted for a defined period.
 - It enables you to define the license terms and conditions that its users must agree to when accessing the data you have to deposit.
 - It can provide either routinely or on request an audit trail of which users access, edit or delete files, including its own actions.
- Will the repository sustain the data's value?
 - Persistent identification: it gives each deposited data collection a publically accessible landing page with a globally unique ID, and (if the metadata is publically accessible) describes how to cite it.
 - Discovery metadata: a landing page identifies the data collection title, creator, and date of deposition.
 - Domain & contextual metadata: supports bi-directional linking of data collection with articles or other related records; metadata can be deposited with the collection as an XML file.
 - Version control: data collections show the date of deposit and most recent change, if any.
 - Data integrity: deposited files are backed up and each copy given a checksum (digital fingerprint).
 - Continuity & succession plan: two copies are held on professionally managed storage at different sites; repository defines how it will ensure continued access if it ceases operation or changes its scope.

- Persistent identification: uses recognized standard schemes (ARK, DOI, PURL, HDL or URN).
- Discovery metadata: complies with relevant standards (Datacite, or OpenAire) so contents are listed in national catalogues or registries.
- Domain & contextual metadata: metadata record for a data collection optionally includes limited domain-specific metadata.
- Version control: depositor can log in to see a change history for their data collection.
- Formats and media: recommends which file formats to use for long-term storage; checks deposited file formats are valid and virus-free.
- Data integrity: performs integrity checks at fixed intervals, monitors file corruption, provides audit record of checks made.
- Continuity & succession plan: two copies minimum are held in different geographic locations.
- Persistent identification: allows a data collection to optionally have multiple identifiers at different levels of granularity.
- Discovery metadata: the landing page for a data collection uses Linked Open Data standards to make metadata machine readable.
- Domain & contextual metadata: offers extensive support for data users to search on domain-specific metadata or controlled vocabulary terms associated with a data collection.
- Open interfaces: integrates with your institution's research information system so that details entered in it do not have to be re-keyed.
- Version control: multiple authorized users may edit a data collection at the same time, and all revisions are recorded.
- Formats and media: commits to migrate data collection to new file format and storage media when existing ones become obsolete.
- Data integrity: performs integrity checks in response to specific events, repairs corrupt data, ensures no one person has write access to all copies.
- Continuity & succession plan: three copies minimum are held in different geographic locations.
- Will the repository support analysis and track data usage?
 - It is clear how you could cite a data collection in the repository, e.g. it describes the title, date, contributors and includes a persistent identifier.
 - Useful search and browse functions are provided.
 - Data collections are accessible for download.
 - Depositor can view access and download statistics.
 - Search and browse functions provided are useful to a researcher in your domain, e.g. by including domain-specific search terms.
 - Data collections are available in open, machine-readable formats.
 - Repository is part of a national or international network providing a common portal for searching data collections.
 - Repository is indexed by a third-party citation tracking service (e.g. Data Citation Index).
 - Free-to-use tools allow data users to extract variables or text terms, or visualize patterns in numeric, image or textual data.
 - Data collections are available in a Linked Open Data format.

The DCC criteria are labeled as a checklist and are not intended to provide a certification outcome. So, while the list is offered from an authoritative source, the lack of a formal certification option kept this from consideration by the USGS Sub-Committee. The DCC questions, however, were found to be relevant and understandable.

National Oceanic and Atmospheric Administration's Unified Framework

This approach originated from the National Oceanic and Atmospheric Administration (NOAA). A paper entitled, "A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Data-sets" (Peng et al. 2015) describes this method. The key components include Preservability, Accessibility, Usability, Production Sustainability, Data Quality Assurance, Data Quality Control/Monitoring, Data Quality Assessment, Transparency/Traceability and Data Integrity. Each of the nine components have rankings to be assigned. The rankings all utilize the scheme below:

- Level 1 Ad Hoc Not Managed
- Level 2 Minimal Managed Limited
- Level 3 Intermediate Managed Defined, Partially Implemented
- Level 4 Advanced Managed Well-Defined, Fully Implemented
- Level 5 Optimal Level 4+ Measured, Controlled, Audit

The NOAA criteria again are fairly comprehensive with an emphasis upon quality elements. The use of a consistent five-level scoring method for each element was straight forward and would be fairly easy to explain. However, there would be some subjectivity in applying the levels across all of the elements and thus, the USGS did not choose this approach. Additionally, the NOAA option is not intended to provide certification.

Data Seal of Approval

The Data Seal of Approval (DANS 2016) approach¹ involved addressing the following criteria:

- The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data, and compliance with disciplinary and ethical norms.
- The data producer provides the data in formats recommended by the data repository.
- The data producer provides the data together with the metadata requested by the data repository.
- The data repository has an explicit mission in the area of digital archiving and promulgates it.
- The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.
- The data repository applies documented processes and procedures for managing data storage.
- The data repository has a plan for long-term preservation of its digital assets.
- Archiving takes place according to explicit work flows across the data life cycle.
- The data repository assumes responsibility from the data producers for access and availability of the digital objects.
- The data repository enables the users to discover and use the data and refer to them in a persistent way.
- The data repository ensures the integrity of the digital objects and the metadata.
- The data repository ensures the authenticity of the digital objects and the metadata.
- The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.
- The data consumer complies with access regulations set by the data repository.
- The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.
- The data consumer respects the applicable licenses of the data repository regarding the use of the data.

Each of the DSA criteria had additional text, statements and questions provided to assist comprehending the information being requested. All of the criteria were judged to be relevant, the length and complexity of the criteria were not considered to be overly burdensome, and the DSA offered a certification option. The USGS had previous experience with the Data Seal of Approval approach. One representative from the Sub-Committee had gone through the Data Seal of Approval process and received certification in 2015. This experience and applicability of the criteria led the USGS to consider adopting this approach.

International Standards Organization

The International Standards Organization (ISO) issued a standard labeled 16363–2012 related to records management. Module 8, entitled, *Becoming A Trusted Digital Repository*, describes the required elements of the ISO standard. The elements are grouped under Organizational Infrastructure, Digital Object Management, Infrastructure and Security Risk Management categories. Additional topics under each element include:

- Organizational Infrastructure
 - Governance and Organizational Viability
 - The repository shall have a mission statement that reflects a commitment to the preservation of, long-term retention of, management of, and access to digital information.

¹ Since October of 2016, the Data Seal of Approval is transitioning certifications in conjunction with the International Council for Science (ICSU) World Data System.

- The repository shall have a Preservation Strategic Plan that defines the approach the repository will take in the long-term support of its mission.
- The repository shall have a Collection Policy or other document that specifies the type of information it will preserve, retain, manage, and provide access to.
- Organizational Structure and Staffing
 - The repository shall have identified and established the duties that it needs to perform and shall have appointed staff with adequate skills and expertise to fulfill those duties.
- Procedural Accountability and Preservation Policy Framework
 - The repository shall have defined its Designated Community and associated knowledge base(s) and shall have these definitions appropriately accessible.
 - The repository shall have Preservation Policies in place to ensure its Preservation Strategic Plan will be met.
 - The repository shall have a documented history of the changes to its operation, procedures, software, and hardware.
 - The repository shall commit to transparency and accountability in all actions supporting the operation and management of the repository that affect the preservation of digital content over time.
 - The repository shall define, collect, track, and appropriately provide its information integrity measurements.
 - The repository shall commit to a regular schedule of self-assessment and external certification.
- Financial Sustainability
 - The repository shall have short- and long-term business planning processes in place to sustain the repository over time.
 - The repository shall have financial practices and procedures which are transparent, compliant with relevant accounting standards and practices, and audited by third parties in accordance with territorial legal requirements.
 - The repository shall have an ongoing commitment to analyze and report on financial risk, benefit, investment, and expenditure (including assets, licenses, and liabilities).
- Contacts, Licenses, and Liabilities
 - The repository shall have and maintain appropriate contracts or deposit agreements for digital materials that it manages, preserves, and/or to which it provides access.
 - The repository shall have contracts or deposit agreements which specify and transfer all necessary preservation rights, and those rights transferred shall be documented.
 - The repository shall have specified all appropriate aspects of acquisition, maintenance, access, and withdrawal in written agreements with depositors and other relevant parties.
 - The repository shall have written policies that indicate when it accepts preservation responsibility for contents of each set of submitted data objects.
 - The repository shall have policies in place to address liabilities and challenges to ownership/rights.
 - The repository shall track and manage intellectual property rights and restrictions on use of repository content as required by deposit agreement, contract, or license.
- Digital Object Management
 - Ingest: Acquisition of Content
 - The repository shall identify the Content Information and the Information Properties that the repository will preserve.
 - The repository shall clearly specify the information that needs to be associated with specific Content Information at the time of its deposit.
 - The repository shall have adequate specifications enabling recognition and parsing of the SIPs [Submission Information Package].
 - The repository shall have mechanisms to appropriately verify the identity of the Producer of all materials.
 - The repository shall have an ingest process which verifies each SIP for completeness and correctness.
 - The repository shall obtain sufficient control over the Digital Objects to preserve them.

- The repository shall provide the producer/depositor with appropriate responses at agreed points during the ingest process.
- The repository shall have contemporaneous records of actions and administration processes that are relevant to content acquisitions.
- Ingest: Creation of the AIP [Archival Information Package]
 - The repository shall have for each AIP or class of AIPs preserved by the repository an associated definition that is adequate for parsing the AIP and fit for long-term preservation needs.
 - The repository shall have a description of how AIPs are constructed from SIPs.
 - The repository shall document the final disposition of all SIPs.
 - The repository shall have and use a convention that generates persistent, unique identifiers for all AIPs.
 - The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains.
 - The repository shall have documented processes for acquiring Preservation Description Information (PDI) for its associated Content Information and acquire PDI in accordance with the documented processes.
 - The repository shall ensure that the Content Information of the AIPs is understandable for their Designated Community at the time of creation of the AIP.
 - The repository shall verify each AIP for completeness and correctness at the point it is created.
 - The repository shall provide an independent mechanism for verifying the integrity of the repository collection/content.
 - The repository shall have contemporaneous records of actions and administration processes that are relevant to AIP creation.
- Preservation Planning
 - The repository shall have documented preservation strategies relevant to its holdings.
 - The repository shall have mechanisms in place for monitoring its preservation environment.
 - The repository shall have mechanisms to change its preservation plans as a result of its monitoring activities.
 - The repository shall provide evidence of the effectiveness of its preservation activities.
- AIP Preservation
 - The repository shall have specifications for how the AIPs are stored down to the bit level.
 - The repository shall have contemporaneous records of actions and administration processes that are relevant to the storage and preservation of the AIPs.
- Information Management
 - The repository shall specify minimum information requirements to enable the Designated Community to discover and identify material of interest.
 - The repository shall capture or create minimum descriptive information and ensure that it is associated with the AIP.
 - The repository shall maintain bi-directional linkage between each AIP and its descriptive information.
- Access Management
 - The repository shall comply with Access Policies.
 - The repository shall follow policies and procedures that enable the dissemination of digital objects that are traceable to the originals, with evidence supporting their authenticity.
- Infrastructure and Security Risk Management
 - Technical Infrastructure Risk Management
 - The repository shall identify and manage the risks to its preservation operations and goals associated with system infrastructure.
 - The repository shall employ technology watches or other technology monitoring notification systems.
 - The repository shall have adequate hardware and software support for backup functionality sufficient for preserving the repository content and tracking repository functions.
 - The repository shall have effective mechanisms to detect bit corruption or loss.

- The repository shall have a process to record and react to the availability of new security updates based on a risk-benefit assessment.
- The repository shall have defined processes for storage media and/or hardware change (e.g., refreshing, migration).
- The repository shall have identified and documented critical processes that affect its ability to comply with mandatory responsibilities.
- The repository shall manage the number and location of copies of all digital objects.
- Security Risk Management
 - The repository shall maintain a systematic analysis of security risk factors associated with data, systems, personnel, and physical plant.
 - The repository shall have implemented controls to adequately address each of the defined security risks.
 - The repository staff shall have delineated roles, responsibilities, and authorizations related to implementing changes within the system.
 - The repository shall have suitable written disaster preparedness and recovery plan(s), including at least one off-site backup of all preserved information together with an offsite copy of the recovery plan(s).

The ISO approach is very complete and exhaustive, as expected from this authoritative organization. Similar to other approaches outlined above, the level of detail sought along with the anticipated time commitment led the USGS to observe this approach would be too burdensome at this stage in the agency's need to establish a process and criteria leading to Trusted Digital Repositories. The ISO approach does provide certification, though.

National Digital Stewardship Alliance

The Library of Congress sponsored the National Digital Stewardship Alliance (NDSA) (Phillips et al. 2013), which developed a "...tiered set of recommendations for how organizations should begin to build or enhance their digital preservation activities. It allows institutions to assess the level of preservation achieved for specific materials in their custody, or their entire preservation infrastructure. It is not designed to assess the robustness of digital preservation programs as a whole since it does not cover such things as policies, staffing, or organizational support. The guidelines are organized into five functional areas that are at the heart of digital preservation systems: storage and geographic location, file fixity and data integrity, information security, metadata, and file formats".

The USGS Data Preservation Sub-Committee built upon the NDSA recommendations and replaced some text (e.g. changed *fixity* to *checksums*) to be more understandable to USGS personnel. The USGS added the element of Physical Media because of the large role that media selections can have on the preservation of agency science data. The levels, beginning with the most elementary achievement of level one, extend to the desired and more demanding level four attainments. These recommendations are presented in **Table 1**.

The NDSA criteria were not intended to provide a certification. The elements and progression are intended more to identify specific areas that digital repositories need to address. Utilizing the table elements could easily be incorporated into an organization's certification submission. As such, the USGS found the approach useful and advocated its use within the agency, but not as the tool recommended for agency certification.

Reviewing the different approaches allowed the USGS to identify the elements included by each. **Table 2** illustrates how the various approaches compare in terms of the elements addressed.

The amount of perceived effort to implement one of the approaches was key to USGS' review. **Table 3** details the perceived level of effort USGS would need to adopt an approach.

Results

Prior to 2015, the USGS had not had an agency-wide plan or strategy to securely preserve its science records. With the organization widely dispersed across the United States, different approaches had been implemented. A consistent, authoritative means was sought to address this fragmented methodology. In addition, increased desires for federal agencies dealing with science records to have documented plans and procedures in place was being noted. To address those needs and after careful evaluation of the pros and

ELEMENT	LEVEL ONE	LEVEL TWO	LEVEL THREE	LEVEL FOUR
Storage and Geographic Location	<ul style="list-style-type: none"> – two complete copies stored physically separate from each other – Transfer the digital content from temporary media into an established storage system – Managed storage system in place 	<ul style="list-style-type: none"> – three complete copies – At least one copy in a different geographic location (offsite locations must follow NARA 1571 guidelines (NARA 2002)) – Document the storage system and storage media 	<ul style="list-style-type: none"> – At least one copy in a geographic location with a different disaster threat (e.g. hurricane area versus an earthquake area) – Maintain an obsolescence monitoring process for the storage system and media 	<ul style="list-style-type: none"> – At least three copies in geographic locations with different disaster threats – Implement a comprehensive plan that keeps files and metadata on currently accessible systems and media
Data Integrity	<ul style="list-style-type: none"> – Verify checksums on ingest, if provided- Create checksums if not provided- Virus check all content 	<ul style="list-style-type: none"> – Verify checksums on all data ingest – Use read only procedures when working with original media 	<ul style="list-style-type: none"> – Verify checksums at fixed intervals – Maintain logs of checksums and supply audit information on demand – Maintain procedures to detect corrupt data – Virus check all content 	<ul style="list-style-type: none"> – Verify checksums of all content in response to specific events or activities – Maintain procedures to replace or repair corrupted data – Ensure no one person has write access to all copies – Create, store, and verify a second, different checksum for all content
Information Security	<ul style="list-style-type: none"> – Identify who has authorization to read, write, move, and delete individual files – Limit authorizations to individual files 	<ul style="list-style-type: none"> – Document access restrictions for content 	<ul style="list-style-type: none"> – Maintain logs of who performed what actions on files, including deletions and preservation actions 	<ul style="list-style-type: none"> – Perform audit of logs
Metadata	<ul style="list-style-type: none"> – Inventory of content and its storage location – Ensure backup and physical separation of inventory information – Adhere to current USGS metadata standards 	<ul style="list-style-type: none"> – Store all relevant database management information – Store information describing changes to the structure or format of the data, including time of occurrence – Provide access to all forms of the metadata 	<ul style="list-style-type: none"> – Preserve standard technical, descriptive, and preservation metadata 	<ul style="list-style-type: none"> – Same as Level 3
File Formats	<ul style="list-style-type: none"> – Encourage the use of a limited set of documented and open file formats, codecs, compression schemes, and encapsulation schemes 	<ul style="list-style-type: none"> – Inventory the file formats in use 	<ul style="list-style-type: none"> – Monitor file format obsolescence issues 	<ul style="list-style-type: none"> – Perform format migrations, emulations (a virtual instance of a previous operating system or procedure) and similar activities
Physical Media	<ul style="list-style-type: none"> – Inventory all physical media utilized including hard disks. 	<ul style="list-style-type: none"> – Develop a plan to utilize trade studies to evaluate medias suitable for USGS purposes. – Begin to transition away from all media utilized that are 10 years or more in age. 	<ul style="list-style-type: none"> – All non-recommended media have been properly disposed of following transition activities. 	<ul style="list-style-type: none"> – Base all media choices on trade studies. – All information is migrated from an older media to a newer media every 3 to 5 years including hard disks.

Table 1: USGS modified *Levels of Digital Preservation*. Derived from Library of Congress, National Digital Stewardship Alliance: Version 1, February 2013. [NARA, National Archives and Records Administration].

	FMM	DCC	UFMS	DSA	ISO 16363	FSPAC/NDSA	DSA/WDS
Organizational Structure, Mandate, Financial Resources	X			X	X		X
Governance Policies	X			X	X		X
Lifecycle Management, Preservation	X	X	X	X	X	X	X
Accessibility	X	X	X	X	X		X
Quality Assurances	X	X	X	X	X	X	
Risk Management	X			X	X		X
Metadata	X	X	X	X	X	X	X
Certification Offered				X	X		X

Table 2: The table matrix shows the criteria established by other organizations to evaluate. The cells indicate the categories addressed by the various criteria examined.

Table source abbreviations used:

FMM = Federal Maturity Model.

DCC = Digital Curation Centre (United Kingdom).

UFMS = A Unified Framework for Measuring Stewardship (NOAA).

DSA = Data Seal of Approval.

ISO 16363 = International Standards Organization No. 16363.

FSPAC/NDSA = USGS Fundamental Science Practices Advisory Council/National Digital Stewardship Alliance.

DCC and DPE = Digital Curation Centre and Digital Preservation Europe.

DSA/WDS = Data Seal of Approval/World Data System (DSA-WDS 2016).

FMM	DCC	UFMS	DSA	ISO 16363	FSPAC/NDSA	DSA/WDS
Medium	Low	Medium	Low	High	Low	Low

Table 3: A qualitative illustration showing perceived effort needed to implement the criteria sets for USGS use. Both cost and complexity were included in the evaluation. The four criteria sets were deemed to have low implementation costs, requiring minimal resources.

cons of the various approaches used by other organizations to review trusted digital repositories, the Data Preservation Sub-Committee recommended using the Data Seal of Approval/World Data System approach to evaluate USGS trusted digital repositories. The ISO approach, while arguably, the most comprehensive, is beyond anticipated USGS resources to implement. Providing a certification path, the anticipated cost, complexity, and usability from the DSA-WDS approach all align well to USGS needs and capabilities. Also, this particular approach suited the requirement for a process that would be viewed as transparent as the DSA-WDS criteria are publically viewable and their reviews are conducted by a blind panel. This approach is laid out in a straight-forward manner allowing the technique to be easily understood by agency staff. Using an internationally recognized approach, one that is drawn from two authoritative bodies, allows the USGS to defend the choice to both USGS staff as well as those outside of the USGS.

The relatively recent, combined DSA-WDS draft criteria were reviewed in February 2016. The 16 primary elements in this approach include addressing the following statements:

- The repository has an explicit mission to provide access to and preserve data in its domain.
- The repository maintains all applicable licenses covering data access and use and monitors compliance.
- The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
- The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
- The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

- The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).
- The repository guarantees the integrity and authenticity of the data.
- The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
- The repository applies documented processes and procedures in managing archival storage of the data.
- The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.
- The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
- Archiving takes place according to defined workflows from ingest to dissemination.
- The repository enables users to discover the data and refer to them in a persistent way through proper citation.
- The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.
- The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
- The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Summary

To address the recent federal mandates requiring U.S. federal agencies to expose and enhance access to federally funded scientific research, the USGS established a cross-agency team to develop a strategic approach for evaluating internal trusted digital repositories for managing scientific assets produced by USGS researchers. The review of multiple criteria developed by national and international organizations to evaluate digital repositories revealed the Data Seal of Approval/World Data System approach as the best option for use by the USGS. This approach will enable the USGS to ensure its repositories are robust and reliable, enabling exposure and access to USGS assets by researchers and the public.

Since 2013, several new data management policies have been developed and implemented by the USGS to preserve and enhance access to scientific assets. The establishment of criteria enabling the certification of agency *Trusted Digital Repositories* was an important element to ensure the preserved digital assets are well managed in reliable systems. The adoption of DSA-WDS Partnership Working Group Catalogue of Common Requirements for trusted digital repository evaluation enhances the lifecycle approach the USGS has adopted to create, maintain, make accessible and preserve its scientific endeavors.²

Acknowledgements

The author would like to thank Keith Kirk, Keith Richmond, Clara Brown, Natalie Latysh, Tara Bell, and Sandra Cooper for their contributions.

Competing Interests

The author has no competing interests to declare.

References

- Data Archiving and Networked Services** 2016 Data Seal of Approval: On-line assessment tool. Netherlands. Available at: <http://www.datasealofapproval.org/en/information/guidelines/> [Last accessed 29 April 2016].
- Data Seal of Approval-World Data System** 2016 DSA-WDS Partnership Working Group Catalogue of Common Requirements. Research Data Alliance. Available at: <https://rd-alliance.org/groups/repository-audit-and-certification-dsa-wds-partnership-wg.html> [Last accessed 29 April 2016].


² The USGS implementation of the criteria to judge Trusted Digital Repositories is expected to proceed for some time in the future. It may also evolve as lessons are learned.

- International Standards Organization** 2012 ISO 16363-2012: AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES. Available at: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=56510 [Last accessed 29 April 2016].
- Joint Working Group of the Federal Records Council and National Archives and Records Administration** 2014 Federal RIM Program Maturity Model User's Guide. Available at: <https://www.archives.gov/records-mgmt/prmd.html> [Last accessed 29 April 2016].
- National Archives and Records Administration** 2002 Archival Storage Standards. Available at: <https://www.archives.gov/files/foia/directives/nara1571.pdf> [Last accessed 29 December 2016].
- Office of Management and Budget** 2013 Open Data Policy – Managing Information as an Asset. Washington, DC. Available at: <https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf> [Last accessed 29 April 2016].
- Office of Science and Technology Policy** 2013 Increasing Access to the Results of Federally Funded Scientific Research. Washington, DC. Available at: https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [Last accessed 29 April 2016].
- Peng, G, Privette, J, Kearns, E, Ritchey, N and Ansari, S** 2015 A UNIFIED FRAMEWORK FOR MEASURING STEWARDSHIP PRACTICES APPLIED TO DIGITAL ENVIRONMENTAL DATASETS. In *Data Science Journal*, Volume 13, 2 February 2015. Available at: <http://datascience.codata.org/articles/abstract/10.2481/dsj.14-049/> [Last accessed 29 April 2016]. DOI: <https://doi.org/10.2481/dsj.14-049>
- Phillips, M, Bailey, J, Goethals, A and Owens, T** 2013 The NDSA Levels of Digital Preservation: An Explanation and Uses. Available at: http://nds.org/documents/NDSA_Levels_Archiving_2013.pdf [Last accessed 2 May 2016].
- Whyte, A** 2015 Where to keep research data: DCC Checklist for Evaluating Data Repositories. v.1 Edinburgh: Digital Curation Centre. Available at: <http://www.dcc.ac.uk/resources/how-guides> [Last accessed 29 April 2016].

How to cite this article: Faundeen, J 2017 Developing Criteria to Establish Trusted Digital Repositories. *Data Science Journal*, 16: 22, pp. 1–13, DOI: <https://doi.org/10.5334/dsj-2017-022>

Submitted: 30 September 2016 **Accepted:** 31 March 2017 **Published:** 19 April 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 