Preservation Storage Criteria: Ongoing Work

September 2018

What are we talking about today?

Preservation Storage Criteria – an ongoing effort by a small group to collect and articulate a set of design attributes that can be used in consideration of Preservation Storage solutions. They are intended as considerations; they are not a set of requirements or a standard.

What is Preservation Storage? Preservation Storage supports digital preservation (*the series of managed activities necessary to ensure continued access to digital materials for as long as necessary* –

Digital Preservation Coalition)

Why Are the Criteria Needed? All digital preservation activities rely on storage, yet there are currently no community guidelines available to aid storage selection. The criteria are intended to help bridge the gap.

Background:

- Originally developed for an iPres2016 workshop called "What is Preservation Storage?"
- Version 3.0 is under development by small group of individuals from the community, with continuing outreach to get feedback at conferences and from the community.

Who is the intended audience?

- **Consumers** of Preservation Storage solutions; and
- **Providers** of Preservation Storage solutions.

Examples of uses:

- Evaluating and comparing Preservation Storage solutions
- Determining gap areas in existing Preservation Storage implementations
- Informing more detailed requirements for Preservation Storage
- As a component of instructional materials on digital preservation
- To help in discussions with IT and other relevant parts of an organization about Preservation Storage
- To help in discussions within the digital preservation field on Preservation Storage

Guiding Principles

- The Criteria should describe characteristics of preservation storage relevant to a wide range of different kinds of institutions with responsibility for preserving digital material, and to organizations providing preservation storage services to other institutions.

- The Criteria omits text that assumes specific architecture, technology, media, content, policy or vendor choices.

- Not all of the Criteria will be applicable to all institutions.

Guiding Principles continued

- The Criteria are not intended to be detailed enough to use as a preservation storage requirements document.
- The Criteria are meant to be used as a foundation for informing preservation storage and to be combined with local policies, applicable regulations, needs and preferences.
- The Criteria do not cover any additional infrastructure needed in combination with preservation storage, e.g. staging areas, testing infrastructure, delivery and management servers.

What else should users take into consideration?

- Institutional requirements, practices, policies
- Applicable regulations, laws
- Areas such as:
  - Confidentiality and privacy
  - Access
  - Risk management framework
  - Financial framework / Cost considerations
  - Business continuity

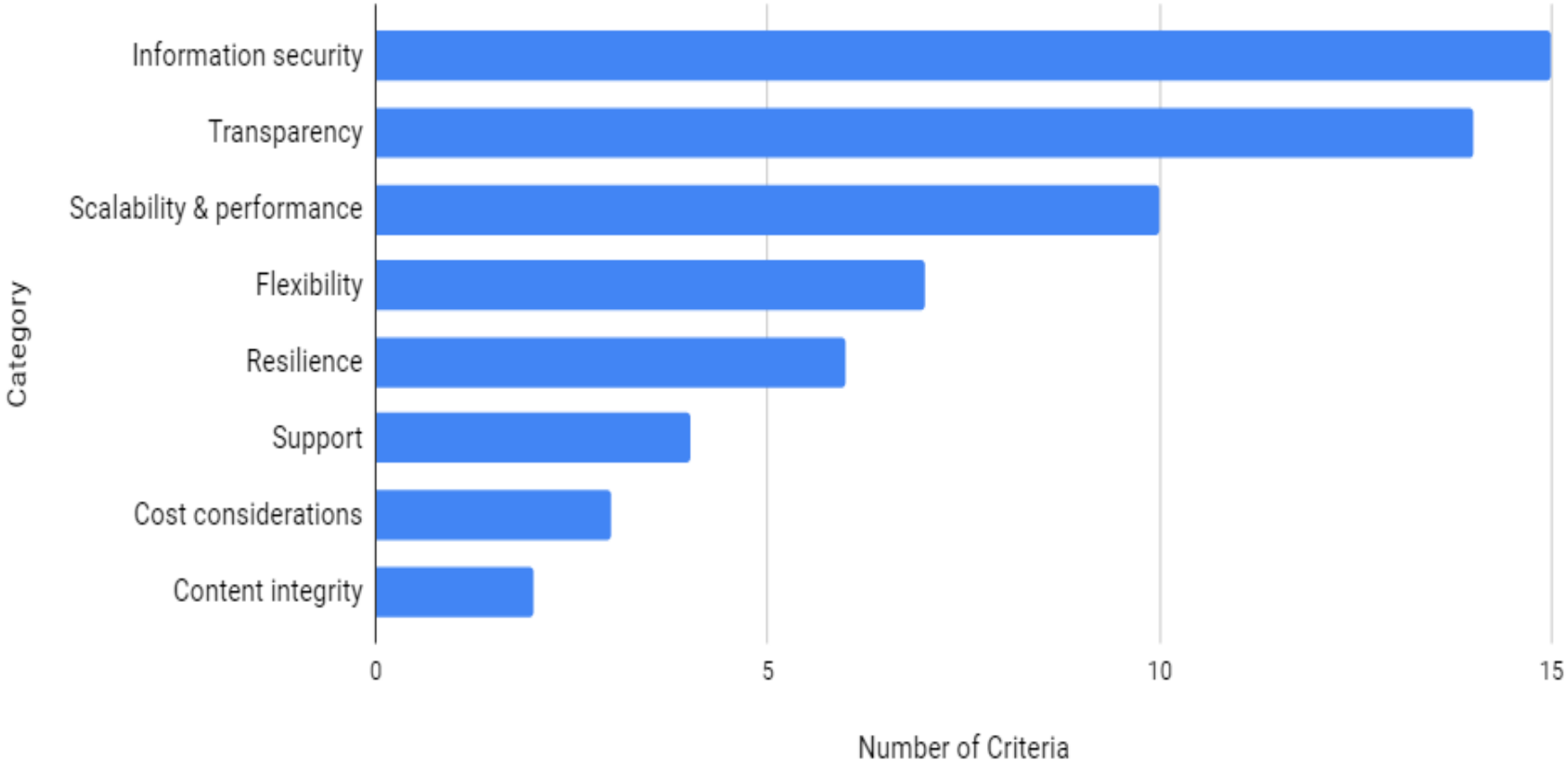
How are the criteria presented?

- Structured list of 61 criteria, each with short description and reference/citation as available or appropriate.
- Each criterion is assigned to one of eight (8) categories for ease of use.
- Accompanied by Criteria Usage Guide with supporting materials on risk, cost, and independence factors.

Criteria Categories

- Content integrity
- Cost considerations
- Flexibility
- Information security
- Resilience
- Scalability & performance
- Support
- Transparency

Criteria by Category



Use Cases: How Are the Criteria Being Used?

- Use criteria as input to discussions within an institution (e.g., between digital curators and IT) or with a storage provider.
- Use criteria to help prioritize within an institution.
- Use criteria to help establish goals for the future.
- Use criteria to help compare services and service providers.
- Use criteria to help compare characteristics of how multiple digital content copies are managed.

Example of How Criteria Might be Used in Use Cases

- http://bit.ly/2Pi3LyN for more examples.

Category	Criterion	Relevant – Y or N?	Provider 1	Provider 2
Content Integrity	Integrity Checking			
Content Integrity	Independent Integrity Checking			
Cost Considerations	Cost-efficient			
Cost Considerations	Energy-efficient			
etc	etc			

Category	Criterion	Description
Content Integrity	Integrity Checking	Performs verifiable and/or auditable checks to detect changes or loss in or across copies (e.g., checksum recalculation, fixity checking, identifying missing files)
Content Integrity	Independent Integrity Checking	Supports fixity checking by other parties (e.g., the content owning institution)
Cost Considerations	Cost-efficient	Costs relatively less overall than other comparable solutions, by being designed with cost efficiencies, for example, has resource pooling and sharing, multi-tenancy (multiple users share the same applications)
Cost Considerations	Energy-efficient	Takes advantage of energy conservation principles and techniques in full or in part. For example, requires less cooling, consumes less power, uses less rack space, as in green computing initiatives.
Cost Considerations	Storage weight	Meets relevant requirements for physical weight as documented in SLA. For example, weight may need to be under a certain amount for a specific floor.

Category	Criterion	Description
Flexibility	Adapts to requirements	Able to adjust storage infrastructure in response to changing local requirements, for example legal requirements or audit results
Flexibility	Constrain location	Enables specification of location (e.g., by geographic region or geopolitical characteristics)
Flexibility	Customizable replication	Supports user-defined replication rules, for example, fewer copies of a specific stream of content
Flexibility	Interoperability	Includes storage components that can be easily integrated with other systems and applications (i.e. plug and play), for example uses standard file access protocols and file system semantics such as NFS, SMB, Rest API's
Flexibility	Open Source	Includes storage components that can be integrated with open source tools, systems, and services in accordance with organization's preferences.

Category	Criterion	Description
Flexibility	Replaceability	Separates storage layer from other systems in the digital preservation environment so that it could be independently refreshed or replaced without affecting the entire infrastructure
Flexibility	Serviceability	Allows for storage maintenance and changes over time without disruption to availability
Information security	Access controls	Provides role-based, access controls for storage infrastructure, e.g. user, staff, admin, to ensure only the appropriate people have the appropriate levels of access
Information security	At-rest server-side encryption with managed keys	Provides encryption, if required, at the storage layer, with no keys for customers to manage
Information security	At-rest server-side encryption with self-managing keys	Provides encryption, if required, at the storage layer, but customers manage encryption keys
Information security	Authentication integration	Integrates relevant organizational authentication systems to authenticate internal and external users of the system.
Information security	Encrypted transfer	Uses an appropriate transport layer encryption at all times when moving content

Category	Criterion	Description
Information security	Geographical independence	Stores multiple redundant copies in geographically-separate locations, at sufficient distances apart, that are not prone to the same natural and human-made disasters and risks
Information security	Multi-tenancy	Supports separate roles/rules/access controls for separate agencies/departments/colleges/faculties etc
Information security	Organizational independence	Manages copies under different organizations, preventing any single organization or individual from causing risk to all copies of the content
Information security	Permanent deletion	Supports requisite deletion by authorized users, in accordance with local policies and rules, in a way that prevents deleted files from being recovered
Information security	Replication	Has documented ability to create redundant, distributed copies of content in reasonable timeframes
Information security	Security protocols	Includes protective measures, controls,and documented procedures to prevent security incidents related to hardware, software, personnel, and physical structures, areas and devices.

Category	Criterion	Description
Information security	System error reporting	Provides immutable logs and/or reports that show all system errors, failures and other critical system activities
Information security	Technical independence	Stores individual copies in different technical solutions (platforms, software including operating systems, hardware, configurations) to prevent all copies from being harmed for example by malware, bugs, or other weaknesses associated with a particular technology.
Information security	Virus/malware detection	Includes software that regularly runs virus checks and malware detection.
Information security	Virus/malware remediation	Provides remediation actions for content with viruses and/or malware, e.g. quarantine, notification, etc.
Resilience	Diverse storage media types	Uses different storage media types / configurations / providers together so that desired levels of independence can be achieved
Resilience	Durable media	Provides documented and acceptable longevity, failure rates, and technical characteristics of the storage media components
Resilience	Error control	Performs error detection and correction 24/7/365 (e.g. using RAID, Erasure coding, ZFS, triple copies/rebuild)

Category	Criterion	Description
Resilience	High availability	Has a high percentage of uptime, i.e. operational for a long length of time, due to techniques such as eliminating single points of failure by having redundant equipment, load-balanced systems and effective monitoring to detect software or hardware failures
Resilience	High resilience	Adapts under stress or faults (e.g. resilient to equipment failures, power outages, attacks, surges in user demand)
Scalability & performance	Recovery and repair	Reviews and replaces or repairs missing or corrupt files in acceptable time frames, in a manner that does not propagate errors; or provides ability and tools to perform these actions independently, e.g. by the content-owning institution
Scalability & performance	Complete exports	Supports the bulk exporting of content and metadata for any reason, at an acceptable rate, for example, as part of an exit strategy
Scalability & performance	Compute power	Meets specified/negotiated computing power for the system or service as documented in the SLA

Category	Criterion	Description
Scalability & performance	Delivery	Meets expectations for delivery from the storage layer, e.g. at a reasonable/negotiated rate and supporting concurrent users
Scalability & performance	File system limits	Able to support long file, path or directory names; large amount of files in a directory, and diverse character encodings
Scalability & performance	I/O performance	Meets specified/negotiated input/output performance levels for the system or service as documented in the SLA
Scalability & performance	Multiple storage tiers	Supports use of multiple storage tiers with different availability levels, e.g. on-line, near-line, off-line
Scalability & performance	Scalable to large data sizes	Able to support very large amounts of content, in terms of number and size of files, and overall volume
Scalability & performance	Supports expansion	Can increase storage capacity over time as needed in accordance with any SLAs
Scalability & performance	Supports reduction	Can decrease storage over time to support deaccessions, transfer of ownership, etc
Scalability & performance	Tiered performance	Meets specified/negotiated performance levels appropriate to material being stored, e.g. Tier1 storage for metadata indexing and searching, Tier2 for caching, Tier3 or lower for bulk storage.

Category	Criterion	Description
Support	Accessibility	Ensures people with disabilities equivalent access to reports, documentation and other content
Support	Independent preservation services	Supports digital preservation services (e.g. migration and transformations with auditable results) by other parties or external tools
Support	Support commitment	Documents commitment to support storage infrastructure, e.g. through SLAs (addressing for example responsibilities, data assurance, response times, end-of-service exit provisions, etc.)
Support	Training	Provides requisite training to appropriate staff across all relevant operational and maintenance tasks
Transparency	Activity monitoring	Supports ability to observe or check activity in the storage infrastructure (e.g. see activity in real-time, examine logs, observe the performance status, determine the overall status or drill-down into activities)
Transparency	Activity reporting	Provides reports about activity in the storage infrastructure (e.g. fixity or virus results, corruption, replacement with good copies)
Transparency	Allow audits	Support independent audits of storage infrastructure and practices in accordance with the SLA

Category	Criterion	Description
Transparency	Assessment information	Provides information needed to support assessments, certifications, audits, and other business activities through for example, documentation, reports, or walkthroughs
Transparency	Content reporting	Provides reports about content in the storage infrastructure (e.g. number of objects/files/formats, average file size, types of objects, size of storage in use)
Transparency	Custom reporting	Supports custom (for example configurable and/or on-demand) reporting of content or activity in the storage infrastructure
Transparency	Data error notification	Notifies content-owners of all data errors, remediation actions and issues in reasonable/expected/negotiated timeframes
Transparency	Documented access	Provides immutable logs and/or reports that show all system access
Transparency	Documented infrastructure	Provides full, complete, current, and available documentation of key processes, services, systems, procedures, known limitations and functions

Category	Criterion	Description
Transparency	Documented provenance	Documents audit/provenance information about all changes, for example about integrity check failures, deletions, modifications, additions, preservation actions; and who or what performed the actions
Transparency	Expose location	Exposes the specific storage location of data to meet SLA requirements
Transparency	Management across storage tiers	Supports management and monitoring across multiple storage availability levels, e.g. on-line, near-line, off-line
Transparency	Open storage formats	Supports open, standard, non-proprietary storage formats, e.g. TAR, archive eXchange format (AXF), LTFS
Transparency	Self-healing transparency	Provides content owners with documentation or notification about any automatic correction or change of data to meet SLA requirements
Transparency		

Next Steps

- Continue to develop version 3.0.
- Get comments on draft version 3.0.
- Present draft version 3.0 at iPres2018 with goals:
  - Improve understandability and clarity of draft Criteria and Criteria usage guide materials; and
  - Expand use case documentation.
- Publish “final” 3.0 version.

Getting Involved/Questions/Comments

- Public Site for Criteria at Open Science Framework (OSF): https://osf.io/sjc6u/
- dpstorage Google group: https://groups.google.com/forum/#!forum/dpstorage
- Conferences & Forums
  - iPRES
  - PASIG
  - NDSA Working Groups