

DataPASS Project – Syndicated Storage Replication Platform, Technical Workplan,  
*Schema Business Rules Requirements*

(Version: 12/06/2007, Micah Altman on behalf of Altman, Beecher, Crabtree, Maynard, McGovern. )

I. Description:

The previously approved project proposal and technical work plan describes the overall project in more detail. To summarize: The syndicated storage platform (SSP) will be used in production to replicate the entire data holdings of the partners. It will also be released publicly as an open-source prototype system.

The system will be implemented on top of another storage technology, the strongest candidate being LOCKSS (with IRODS remaining of interest for experimentation) The first deliverable (for the end of Q1 CY08) for the Syndicated Storage Replication Platform is a schema that will define inter-archival replication commitments, and which will be used to drive automated configuration, monitoring, and auditing of the platform, using software developed that will be developed later in the project.

This document proposes a set of business requirements for that *schema*. The full requirements for the platform are not in the scope of this document, however, a number of potential requirements for the syndicated storage platform as a whole have emerged from discussions and we comment on the nature of such requirements and their relationship to the schema.

II. Previously determined requirements for the platform:

Derived from the proposal to LoC:

- The platform must be capable of storing multiple replicas of the content
- The platform must be capable of replicating the entire non-confidential public holdings of quantitative data, metadata, and documentation of each partner. ( Replication of video and audio materials is optional but not required. Data subject to strict confidentiality/enclave requirements may be excluded. )
- The platform must support asymmetric replications resource commitments by the partners. (Different partners will have different sized collections, and contribute different sized storage resources. All holdings will not be replicated at all sites.)
- The platform must support some form of versioning. At minimum, previous versions of holdings deposited into the platform must be retained.
- The platform must support restoration of groups of content to the owning archive.
- The platform must support the restoration and transfer of the entire holdings of a designated archive to another partner, should that archive undergo institutional failure.

III. Proposed schema requirements:

Schema should support configuration, monitoring, and auditing of the following features.

- Replication policies
  - Minimum/desired number of replicas in entire
  - Minimum/desired freshness of replicas in entire

- Minimum/desired frequency of self verification of replicas in entire
- Minimum/desired frequency of external verification of replicas in entire
  
- Versioning, etc.
  - Versioning is required/optional/unavailable for SPP
  - Replacement/Deletion is permitted/forbidden/unavailable for SPP
  
- Resource Commitments
  - Storage/network available per replica
  - Storage/network promised to replica
  
- Auditing/reporting
  - Report on replication/verification history/frequency
  - Report on availability of nodes
  - Report/log any failures
  - Provide location of local audit trail
  - Report on resource commitments asserted/used
  - Report on state of archival unit, locations
  - Report of verification activity
  
- Archival Unit (AU) Level Description
  - Descriptive meta-data for archival unit
  - Harvesting server, protocol, and identifier for AU
  - AU Owner
  - Per-AU Replication Policies \*
  - Per-AU Versioning Policies \*
  - Per-AU Resource commitments \*

*Comment:* AU's are expected to be coarse-grained, and represent entire archives or major divisions within archives' collections. Implemented support for starred items are desirable but not required.

- TRAC
  - Schema should provide a *reference* to external document that describes conformance with TRAC criteria.
  - Document a mapping from schema elements (or groups of elements) to TRAC criteria supported (in whole or in part) by those element.

*Comments: see institutional requirements below.*

The following are desirable, but not required:

- Documentation of other node characteristics
- Archival preferential ordering for node characteristics

- Implementation in a standard policy language
- Mapping from schema to IRODS rules

#### IV. Other proposed requirements questions:

##### A. Institutional requirements

###### A.1 For institutions hosting a node in the SSP

- should sign a *uniform* partnership agreement describing hosting responsibilities
  - *partnership agreement should aspire compliance with identified TRAC criteria (see appendix B) minimal TRAC compliance (?)*
  - *partnership agreement should include authentication, authorization, security, confidentiality terms*
  - *confidentiality terms should include notice that (1) sensitive data should be provided in encrypted form only; (2) encryption keys should be placed in escrow with the partnership, but are not stored in or a part of the SSP itself*
  - *partnership terms will describe the terms and conditions under which partners may subcontract the actual storage infrastructure to another service (such as Amazon E3). This should include, at minimum, a provision that allows the partner direct access to the storage provided by the third-party, in the case of institutional failure of the partner.*
  - *partnership agreement should include other representations or warranties of performance*
- Should provide contact information

Data-PASS is striving toward becoming a virtual organization conforming with preservation standards and practices. As such, it is a long-term goal that the virtual organization as a whole be able to demonstrate conformance with these standards, but not essential that every participating host of the SSP platform be conformant.

###### A.2 For institutions providing content:

- Should represent that they have sufficient authority/rights over content deposit to comply with the terms of use below
- Should agree that content can be disseminated by other Data-PASS partners (subject to documented terms of use for each AU) should depositor suffer an organizational failure
- Should provide contact information

##### B. Content Requirements

###### B.1 Depositor Requirements

- Content placed in the SSP must be deposited under the of an institution qualified under (A.2)

###### B.2 Metadata Requirements

All content should be clearly linked to *descriptive and structural* metadata complying with the existing Data-PASS metadata agreement. In order for content in SSP to ever be disseminated by partners other than the depositor, sufficient metadata must be preserved regarding rights. The rights and terms

associated with content in the storage system will be more complex and varied than those collected under the uniform Data-PASS deposit agreement, thus finer grained rights metadata is required. In addition additional rights metadata categories should be developed and documented in the DDI instance, in support of the terms of A.2:

- Rights metadata should assert the right is given to be replicated, reformatted etc for archival purposes under standard terms. And should reference enabling document.
- Rights metadata should assert right to disseminate *under original terms of deposit*, in the case of archival failure. And should reference enabling document.
- Metadata should indicate click-through terms, and should assert whether click-through terms are sufficient. If click-through terms insufficient, should reference document describing terms of use (not necessarily machine actionable) .
- Metadata should indicate whether content may be disseminated by other partner archives under the given terms immediately, or only in the case of archival institutional failure.

:

### B.3 Content completeness

Current Data-PASS metadata standards require depositor agreements to be bundled with data collections in order to support effective transfer to LC and/or other partners. Content may not be provably legally usable if agreements covering replicated content is lost. The above completeness requirements are intended to address this.

- *All data-pass required metadata, partner agreements, depositor agreements, and any other documents describing/enabling rights/terms of use that are referenced in object metadata as required licensing or rights terms should be replicated in syndicated storage system. (?)*
- *All documents supporting other institutional requirements above should be in syndicated storage system?*

### C. SSP Technology/Operational Requirements

- Cost effectiveness
  - Should be implementable using commodity hardware
  - Should be maintainable in partners' IT environments
- Intellectual Property
  - Should be available under OSS approved Open Source license
  - If any patented technology used, must have royalty-free, perpetual license for use in conjunction with the open-source software
- Harvesting
  - Must be able to harvest public/restricted content from standard LOCKSS-enabled websites.
  - Must be able to harvest public/restricted content from dataverses
- Authorization/Authentication
  - Must be able to authenticate users and enforce roles, responsibilities, sufficient to enforce content access policies
  - Federated authentication (e.g. shibboleth)

- Performance

This is not a system used for on-line delivery, so there are no intensive load requirements, instead the following performance requirements apply:

- Must be able to complete replication of incremental changes nightly.
- Must be able to complete restore request for individual unit overnight.
- Must be able to restore complete contents of a member archive within a week.

## **Appendix A: Virtual Organization Conformance**

(Nancy McGovern)

The Digital Preservation Management curriculum<sup>1</sup> features the definition of five stages of organizational development for digital preservation:

1. Acknowledge: understanding that digital preservation is a local concern
2. Act: initiating digital preservation projects
3. Consolidate: segueing from projects to programs
4. Institutionalize: incorporating the larger environment and rationalizing DP programs
5. Externalize: embracing inter-institutional collaboration and dependency

Data-PASS is striving towards stage 5, effectively a virtual organization composed of the partners. An organization may exhibit stage 5 behavior between stages 2 and 4, but at least one organization must have achieved stage 4 for sustainable stage 5 collaboration to be achievable. By stage 3, organizations commit to aligning with the prevailing standards and practice of the digital preservation community. By stage 4, a repository can demonstrate how it conforms.

In the context of the syndicated storage project, applying the five stages means that the Data-PASS partners do not individually have to be conformant with standards and practice, but cumulatively Data-PASS should be able to demonstrate that it is conformant.

In principle, the standards and practice that Data-PASS is aligned (or aligning) with include:

- the Open Archival Information Systems (OAIS) Reference Model, an ISO standard
- Trusted Digital Repositories: Attributes and Responsibilities, a community guidance document
- Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, a developing ISO standard
- A Framework of Guidance for Building Good Digital Collections, community guidance from NISO
- Preservation Metadata Implementation Strategies (PREMIS) Data Dictionary, a community document that may be an emerging standard.

Demonstrating conformance with these examples of digital preservation community standards and practice entails explicitly documenting the approach of a repository is addressing the requirements (mapping actions and developments to the requirements) and being able to provide evidence that the requirements are being addressed. The TRAC requirements incorporate the essential requirements of both the Trusted Digital Repositories and the OAIS documents.

---

<sup>1</sup> The Digital Preservation Management curriculum was developed by Anne R. Kenney, Interim University Library at Cornell University Library, and Nancy Y. McGovern, Digital Preservation Officer at ICPSR.

**Appendix B: Desirable TRAC Criteria Pertaining to Data-PASS Syndicated Storage Agreements and Actions**

TRAC criteria	Syndicated Storage Response	Notes
<b>A. Organizational Infrastructure</b>		
<b>A1. Governance &amp; organizational viability</b>		
A1.1. Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information.		Should be true for all partners – motivation for syndicated storage
A1.2. Repository has an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.		Pertains to the agreements arrangements needed for syndicated storage partners
<b>A2. Organizational structure &amp; staffing</b>		
A2.1. Repository has identified and established the duties that it needs to perform and has appointed staff with adequate skills and experience to fulfill these duties.		This pertains to the storage manager role – the responsibilities of that role need to be defined for syndicated storage participants.
<b>A3. Procedural accountability &amp; policy framework</b>		
A3.2. Repository has procedures and policies in place, and mechanisms for their review, update, and development as the repository grows and as technology and community practice evolve.		For syndicated storage, relevant polices and procedures would pertain to archival storage, replication, and related issues.
A3.3. Repository maintains written policies that specify the nature of any legal permissions required to preserve digital content over time, and repository can demonstrate that these permissions have been acquired when needed.		Partners need to have documentation of the right to ensure redundant storage for preservation – supporting documentation.
A3.4. Repository is committed to formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements.		This pertains to technological changes that affect the stored content.
A3.6. Repository has a documented history of the changes to its operations, procedures, software, and hardware that, where appropriate, is linked to relevant preservation strategies and describes potential effects on preserving digital content.		This kind of documentation will be needed for syndicated storage.
A3.7. Repository commits to transparency and accountability in all actions supporting the operation		Policies for syndicated storage need to be explicit and accessible.

TRAC criteria	Syndicated Storage Response	Notes
and management of the repository, especially those that affect the preservation of digital content over time		
A3.8 Repository commits to defining, collecting, tracking, and providing, on demand, its information integrity measurements.		Collaborative agreements and means for ensuring integrity of jointly managed content need to be explicit and accessible.
A3.9 Repository commits to a regular schedule of self-assessment and certification and, if certified, commits to notifying certifying bodies of operational changes that will change or nullify its certification status.		Syndicated storage needs to be auditable.
<b>A4. Financial sustainability</b>		
A4.1. Repository has short- and long-term business planning processes in place to sustain the repository over time.		The funding needs to be documented and shown to be sustainable.
A4.5. Repository commits to monitoring for and bridging gaps in funding.		Long-term plans should be in place by the end of NDIIPP funding.
<b>A5. Contracts, Licenses and Liabilities</b>		
A5.1 If repository manages, preserves, and/or provides access to digital materials on behalf of another organization, it has and maintains appropriate contracts or deposit agreements.		For syndicated storage, these would be the syndicated storage agreements rather than individual deposit agreements.
<b>B. Digital Object Management</b>		
<b>B.1 Ingest: acquisition of content</b>		
B1.5. Repository obtains sufficient physical control over the digital objects to preserve them (Ingest: content acquisition).		Chain of custody needs to be extended to syndicated storage.
B1.8. Repository has contemporaneous records of actions and administration processes that are relevant to preservation.		Audit trails in the form of logs and other means needs to be accessible for syndicated storage management.
<b>B.2 Ingest: creation of the archivable package</b>		
B2.5. Repository has and uses a naming convention that generates visible, persistent, unique identifiers for all archived objects (i.e., AIPs).		The means to maintain persistent identifiers need to be in place and auditable for syndicated storage.
B2.7. Repository demonstrates that it has access to necessary tools and resources to establish authoritative semantic or technical context of the digital objects it contains (i.e., access to appropriate international		Syndicated storage will have to determine how to address this requirement for members. This is about redundant storage not about rendering content for other partners.

TRAC criteria	Syndicated Storage Response	Notes
Representation Information and format registries).		
B2.12 Repository provides an independent mechanism for audit of the integrity of the repository collection/content.		This is a question for syndicated storage management.
<b>B.3 Preservation Planning</b>		
B3.2. Repository has mechanisms in place for monitoring and notification when Representation Information (including formats) approaches obsolescence or is no longer viable.		Data-PASS content should be monitored for obsolescence and actions may affect stored content.
<b>B.4 Archival storage &amp; preservation/ maintenance of AIPs</b>		
B4.4 Repository actively monitors integrity of archival objects (i.e., AIPs).		This requirement extends to syndicated storage.
B4.5 Repository has contemporaneous records of actions and administration processes that are relevant to preservation (Archival Storage).		Audit trails for archival storage actions extend to syndicated storage.
<b>B.5 Information Management</b>		
B5.2 Repository captures or creates minimum descriptive metadata and ensures that it is associated with the archived object (i.e., AIP).		Descriptive metadata needs to be able to be associated with AIPs in syndicated storage.
B5.4 Repository can demonstrate that referential integrity is maintained between all archived objects (i.e., AIPs) and associated descriptive information.		Referential integrity needs to be maintained for AIPs in syndicated storage.
<b>B.6 Access Management</b>		
B6.3 Repository ensures that agreements applicable to access conditions are adhered to.		Syndicated storage addresses preservation not access – this is controlled access for shared content storage.
B6.4 Repository has documented and implemented access policies (authorization rules, authentication requirements) consistent with deposit agreements for stored objects.		This pertains to rules for managing stored content – not to individual deposit agreements. Who has access to syndicated storage and what are they able to do / not do?
<b>C. Technologies, Technical Infrastructure &amp; Security</b>		
<b>C1. System Infrastructure</b>		
C1.1 Repository functions on well-supported operating systems and other core infrastructural software.		
C1.2 Repository ensures that it has adequate hardware and software support for backup functionality sufficient for the repository’s services and for the data held, e.g., metadata associated with access controls, repository main content.		

TRAC criteria	Syndicated Storage Response	Notes
C1.3 Repository manages the number and location of copies of all digital objects.		
C1.4 Repository has mechanisms in place to ensure any/multiple copies of digital objects are synchronized.		
C1.5 Repository has effective mechanisms to detect bit corruption or loss.		
C1.6 Repository reports to its administration all incidents of data corruption or loss, and steps taken to repair/replace corrupt or lost data.		
C1.7 Repository has defined processes for storage media and/or hardware change (e.g., refreshing, migration).		
C1.8 Repository has a documented change management process that identifies changes to critical processes that potentially affect the repository's ability to comply with its mandatory responsibilities..		
C1.9 Repository has a process for testing the effect of critical changes to the system.		
C1.10 Repository has a process to react to the availability of new software security updates based on a risk-benefit assessment.		
<b>C.2 Appropriate technologies</b>		
C2.1 Repository has hardware technologies appropriate to the services it provides to its designated communities and has procedures in place to receive and monitor notifications, and evaluate when hardware technology changes are needed.		The focus for syndicated storage is on appropriate technology rather than designated communities.
C2.2 Repository has software technologies appropriate to the services it provides to its designated community(ies) and has procedures in place to receive and monitor notifications, and evaluate when software technology changes are needed.		Same as C.2.1

---