Trusted Digital Repository Design: a Policy-Driven Approach

Chien-Yi Hou; University of North Carolina, Chapel Hill; North Carolina, United States; chienyi@unc.edu
Caryn Wojcik; Records Management Services, State of Michigan; Lansing, Michigan, United States; wojcikc@michigan.gov
Richard Marciano; University of North Carolina, Chapel Hill; North Carolina, United States; richard_marciano@unc.edu

Abstract

The "Distributed Custodial Archival Preservation Environments" project, DCAPE, proposes a framework to support institution-specific preservation policies and to build a distributed production preservation environment that meets the needs of archival repositories for trusted archival preservation services. The project focuses on the development of trusted digital repositories driven by institutional policies that can be highly customized. The DCAPE team is composed of technologists and archivists from state archives, university archives, and cultural institutions. Our approach is based on the emerging ISO/DIS 16363 standard on "Audit and Certification of Trustworthy Digital Repositories." The services proposed follow a particular flavor of microservices and rules, based on an iRODS [1] implementation.

The DCAPE framework is built on a state-of-the-art rule-based data management system. The system provides developers with the mechanism to design customized workflows as rules that will be executed by the internal rule engine. The DCAPE project hides the complexities of designing these rules and provides a user-friendly interface for users to select from the list of pre-defined rules to implement their policies. These policies can be applied at an institution level or at the collection level. In this paper, we will discuss the design of the framework, the preservation workflow, and the interface.

1. Introduction

The goal of the DCAPE project is to build a distributed production preservation environment that meets the needs of archival repositories for trusted archival preservation services for electronic records. The project is built on the philosophy that individual archival repositories may not have the resources and skills necessary to build and maintain an internal digital preservation system. The DCAPE preservation environment is built on top of a trusted digital repository infrastructure that is assembled from a rule-based data management system, commodity storage systems, and sustainable preservation services.

Every institution or archive has its own policies to manage its preservation environment and records. These policies usually include management of archival storage, validation, and trustworthiness. The enforcement of these policies is typically labor-intensive, and the implementation of these policies as automated processes requires strong technical skills. The DCAPE project plans to solve these two problems together by offering a set of services that are deemed to be essential among archivists, and by providing an interface for the archivists to manage these services as their repository's policies. These services are implemented as rules in the preservation environment and these machine-actionable rules are designed to be highly customizable by the archivists through the DCAPE interface.

The DCAPE project is a collaborative effort involving multiple "medium-scaled" preservation communities who share the explicit goal of defining the common set of services that are needed by all participating institutions (state archives, university archives, cultural institutions, etc.). The team is also formulating workflows that will allow each DCAPE user to make choices about the preservation environment that meet their own unique needs. The team conducted an assessment of OAIS capabilities relevant to the project, based on requirements from their own institutions. This led to a specification with close to 100 policies. The team then selected an initial set of 52 policies to focus on for the initial system development. Out of these 52 policies, a working subset was extracted, consisting of approximately 26 rules (see Appendix 1 for details). A research testbed and an SLA (service-level agreement) were also established so that records from the partner institutions could be loaded into a testbed.

2. The DCAPE Framework

The DCAPE Framework is built on top of the iRODS middleware system. iRODS provides the capability to build a distributed preservation environment.it's the iRODS rule engine allows users to customize rules to manage the records. The team separated the DCAPE resources into three different functional areas: *(1) Virtual Loading Dock, (2) Preservation Area, and (3) Reference Room*. Each functional area serves different purposes through the life cycle of the records in the collections. The *Virtual Loading Dock* is used as a staging area for users to manage Submission Information Packages (SIPs). The records that are submitted by the record providers will be examined and cleaned at this phase. The system might ask the providers to re-submit a record if it is infected by virus or corrupted during the transmission. Once a SIP is accepted, DCAPE will generate the corresponding Archival Information Package (AIP) and store it in the Preservation Area. *The Preservation Area* is meant to be used for long-term preservation, and is not accessible to the public. The main task at this stage is to ensure the integrity and stability of the record. Replication for disaster prevention is handled at this stage. If an AIP is approved to be shared with the public, the AIP will be re-packaged as a Dissemination Information Package (DIP) and moved to the Reference Room. The *Reference Room* is an area that could be accessible by the public. Users who are interested in the record will have the opportunity to download the record from the Reference Room. Figure 1. illustrates the framework. V1, P1, and R1 are physical storage resources that are all managed by iRODS.

3. The DCAPE Capabilities and Rules

To build a digital preservation service, one has to decide what capabilities a service needs to support. The DCAPE team initially focused on the implementation of a set of capabilities that were deemed the most essential across the partner archival institutions. The project archivist partners had a series of discussions and defined a set of common capabilities [2] that were desired across the various institutions. These capabilities were stated in "plain English" and handed to the DCAPE rule development team to translate them into machine-actionable iRODS rules. Not all of desired capabilities can be implemented as machine-actionable rules. Some capabilities are functional statements instead of rules. For example, DCAPE capability #24 states, "We need a search interface." Further work will need to be done by the project team to develop a search interface for the Reference Room. Specifically, the team will need to identify data and metadata that will be searchable, and will need to develop a mechanism for conducting the search. Other capabilities can be easily translated into rules. For example, DCAPE capability #5 states, "We need to do a virus check when the record is ingested." A rule for this capability will trigger a virus check procedure to examine the record after the file is uploaded to the environment.

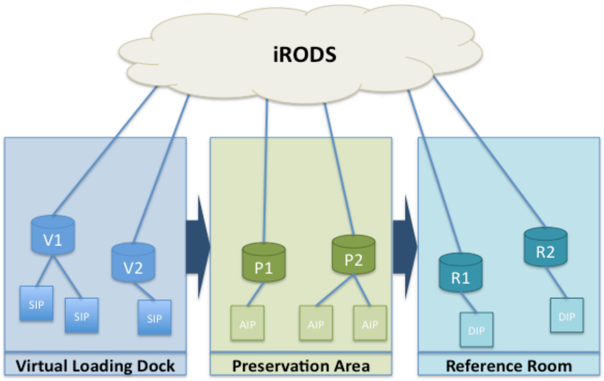


Figure 1. The DCAPE Framework

An iRODS rule has four components (trigger, condition, chain of workflow, and recovery chain). The trigger is an iRODS action definition. It could be an action to upload a file or an action to remove a file. The rule will be triggered if the defined action takes place and the specified condition is also satisfied. Once the rule is triggered, it will execute the chain of the workflow that is composited by micro-services. The chain of workflow could contain multiple micro-services and even other chains of workflows within it. Micro-services are small procedures that perform a certain task. Micro-services could also invoke some external tools, like JHOVE and DROID, to perform some preservation related procedures. If an error occurs during the execution of the workflow, the recovery chain will be performed.

Figure 2. shows some of the initial services the DCAPE team intends to deliver in each functional area. The team will wrap open-source software as micro-services to provide some of these services. For example, the team can use FITS [3] to integrate other tools to do file type validation and metadata extraction. In terms of preservation metadata, iRODS provides a very flexible metadata structure. DCAPE users can add user-defined metadata at the collection or at the object level. If there are certain metadata formats a DCAPE user would like to apply to its records, whether it is PREMIS [4] or METS [5], a template can be pre-defined and a preferred format can be chosen.

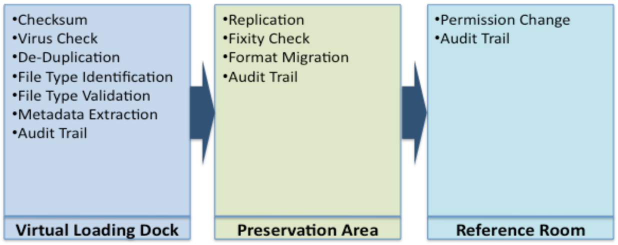


Figure 2. The DCAPE Proposed Services

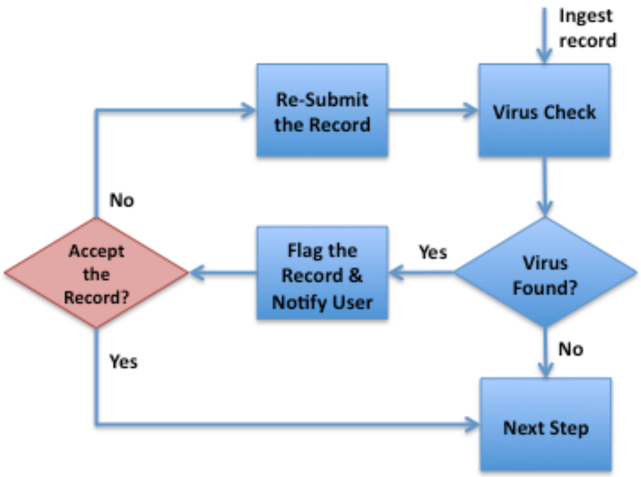


Figure 3. Virus Check Workflow

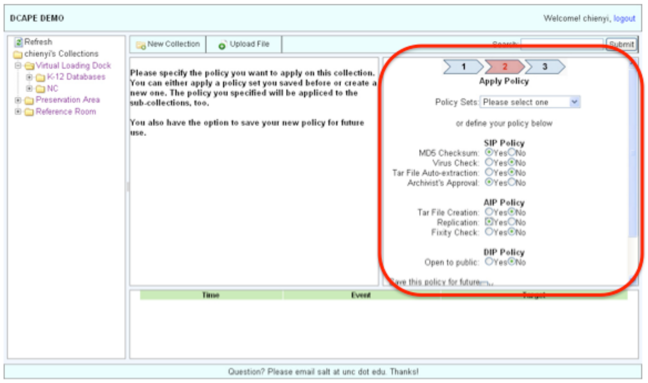
Most of the services are specific for a certain functional area, but there is one service that the project team considered to be essential for all three functional areas. Integrity is an essential component of a trusted digital repository, so all of the functional areas will have an audit trail. The audit trail will provide the evidence of what has taken place in the repository. iRODS supports audit trails and tracks processing at the object level. The DCAPE team is also working on giving users the flexibility to decide which actions they want to track.

Besides the services mentioned in Figure 2., the team will also design some recovery services into the workflow for users to choose when something abnormal takes place. Figure 3. is an example of how to deal with the situation when a virus is detected in a record. Users have the ability to weigh in and decide how they would like this virus-infected record to be treated. If the preference is specified in a pre-defined way, the system will execute the workflow automatically, run the procedure and notify users directly.

4. The DCAPE Interface

The DCAPE partners also identified the need for a customizable interface to manage their repository policies. This is a key aspect of the development, as well as a time-consuming endeavor. The goal is to hide the details of the rules and give the users an easier way to manage their policies. There are some policies that are used by all the institutions, but there are some special policies that are unique to each institution. The goal is to give archivists the flexibility to manage policies at the collection level or even at the repository level. The DCAPE interface provides a policy template for users to select desired policies and associate them with certain collections. Figure 4. is a snapshot of the DCAPE interface prototype. Archivists can use radio buttons to decide whether they want this service to be applied to the collection or not. They can also save policies and apply the saved policy set to another collection in the future.

The DCAPE interface not only provides the ability to customize the preservation workflow, it also shows the archivist the status of the records, and the metadata,. Figures 5. through 6. represent snapshots of these features.



Please specify the policy you want to apply on this collection. You can either apply a policy set you saved before or create a new one. The policy you specified will be appliced to the sub-collections, too.

You also have the option to save your new policy for future use.

Policy Sets: Please select one

or define your policy below

SIP Policy
MD5 Checksum: ● Yes ○ No
Virus Check: ○ Yes ● No
Tar File Auto-extraction: ○ Yes ● No
Archivist's Approval: ● Yes ○ No

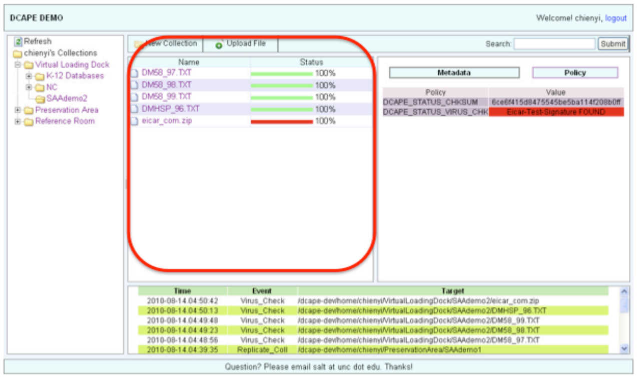
AIP Policy
Tar File Creation: ○ Yes ● No
Replication: ● Yes ○ No
Fixity Check: ○ Yes ● No

DIP Policy
Open to public: ○ Yes ● No

Save this policy for future...

Question? Please email salt at unc dot edu. Thanks!

Figure 4. The DCAPE Interface – Apply Policy



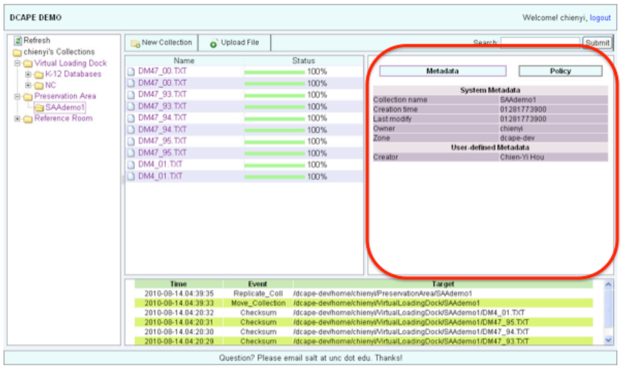
Name	Status
DM58_97.TXT	100%
DM58_98.TXT	100%
DM58_99.TXT	100%
DMHSP_96.TXT	100%
eicar_com.zip	100%

Policy	Value
DCAPE_STATUS_CHKSUM	6ce6f415d8475545be5ba114f208b0ff
DCAPE_STATUS_VIRUS_CHK	Eicar-Test-Signature FOUND

Time	Event	Target
2010-08-14.04:50:42	Virus_Check	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo2/eicar_com.zip
2010-08-14.04:50:13	Virus_Check	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo2/DMHSP_96.TXT
2010-08-14.04:49:48	Virus_Check	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo2/DM58_99.TXT
2010-08-14.04:49:23	Virus_Check	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo2/DM58_98.TXT
2010-08-14.04:48:56	Virus_Check	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo2/DM58_97.TXT
2010-08-14.04:39:35	Replicate_Coll	/dcape-dev/home/chienyi/PreservationArea/SAAdemo1

Question? Please email salt at unc dot edu. Thanks!

Figure 5. The DCAPE Interface – Status and Error Report



Name	Status
DM47_00.TXT	100%
DM47_00.TXT	100%
DM47_93.TXT	100%
DM47_93.TXT	100%
DM47_94.TXT	100%
DM47_94.TXT	100%
DM47_95.TXT	100%
DM47_95.TXT	100%
DM4_01.TXT	100%
DM4_01.TXT	100%

System Metadata	
Collection name	SAAdemo1
Creation time	01281773900
Last modify	01281773900
Owner	chienyi
Zone	dcape-dev
User-defined Metadata	
Creator	Chien-Yi Hou

Time	Event	Target
2010-08-14.04:39:35	Replicate_Coll	/dcape-dev/home/chienyi/PreservationArea/SAAdemo1
2010-08-14.04:39:33	Move_Collection	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo1
2010-08-14.04:20:32	Checksum	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo1/DM4_01.TXT
2010-08-14.04:20:31	Checksum	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo1/DM47_95.TXT
2010-08-14.04:20:30	Checksum	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo1/DM47_94.TXT
2010-08-14.04:20:29	Checksum	/dcape-dev/home/chienyi/VirtualLoadingDock/SAAdemo1/DM47_93.TXT

Question? Please email salt at unc dot edu. Thanks!

Figure 6. The DCAPE Interface – Metadata

5. Summary

A trusted digital preservation service with customizable policies is a very useful and innovative capability for archivists. The DCAPE team has successfully demonstrated a proof of concept and prototype on how to use the DCAPE interface to customize the policies for the repository. The project is currently adding additional services and refining the DCAPE interface to give users an optimized usage experience. The team is also working on developing a business model that will lead to hosted trusted digital repository services.

6. Acknowledgements

This project was funded by NHPRC Records Projects grant NAR08-RE-10010-08, "Distributed Custodial Archival Preservation Environments", 2008-2011. In addition the PoDRI project (Policy-Driven Repository Interoperability), funded through IMLS grant LG-06-09-0184-09, has influenced various aspects of this development.

More information on the DCAPE project including project partners, can be found at http://dcape.org. In particular, we wish to acknowledge the collaboration and parthership of the following institutions and individual, from states: *California* (Chris Garmire), *Kansas* (Scott Leonard), *Kentucky* (Glen McAninch, Mark Myers), *Michigan* (Mark Harvey, Caryn Wojcik), *North Carolina* (Kelley Eubank, Lisa Gregory, Jennifer Ricker, Amy Rudersdorf, Druscie Simpson), *New York* (Michael Martin, Ann Marie Przybyla, Bonita Weddle); universties: *Tufts University* (Veronica Martzahl, Eliot Wilczek), *UNC Chapel Hill* (David Pcolar), *University of Wisconsin* (Kristin Eschenfelder), *West Virginia University* (Donald Adjeroh); and cultural institutions: *Getty Research Institute* (Joe Shubitowski), *Smithsonian Institution Archives* (Ricc Ferrante).

References

[1] iRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. http://www.irods.org

[2] Richard Marciano, Chien-Yi Hou, et al. "Community-Driven Development of Preservation Services", the iRODS User meeting in Chapel Hill, March 24-26, 2010.

[3] FITS: File Information Tool Set. http://code.google.com/p/fits/

[4] PREMIS: http://www.loc.gov/standards/premis/

[5] METS: Metadata Encoding & Transmission Standard. http://www.loc.gov/standards/mets/

Author Biography

Chien-Yi Hou is a research associate at School of Information and Library Science in University of North Carolina, Chapel Hill. He obtained his degree in Computer and Information Science from the National Chiao Tung University in Taiwan and M.S. in Computer Science and Engineering from University of California, San Diego.

Caryn Wojcik is a Government Records Archivist for the State of Michigan where she has worked since 1996. She received both her Bachelor's degree in history (1993) and her Master's degree in information and library science (1995) from the University of Michigan, Ann Arbor. She is a Certified Archivist (received 2001).. She is secretary to the Board of Directors of NAGARA.

Richard Marciano is Director of the Sustainable Archives & Leveraging Technologies (SALT) lab, a Professor in the School of Library and Information Science at the University of North Carolina at Chapel Hill.. Richard holds degrees in Avionics and Electrical Engineering; M.S. and Ph.D. in Computer Science from the University of Iowa, and worked as a Postdoc in Computational Geography.

Appendix 1

Mapping from 52 “OAIS Criteria” to “ISO/DIS 16363 Items” to “Machine-Actionable Rules” to an initial set of 26 “DCAPE Items”

ISO/DIS 16363 Items are referenced from:
http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/CombinedMetricsDocumentsFollowingFaceToFace

OAIS Criteria	ISO/DIS 16363 Items	DCAPE Machine-Actionable Rules	DCAPE Items
1. Address liability and challenges to ownership/rights.	A3.2.2 A5.1.3 A5.1.4 A5.2	Map from submission template to access and distribution controls	
2. Identify the content information and the information properties that the repository will preserve.	B1.1 B1.1.2	Define templates that specify required metadata and parameters for rules that are required to enforce properties	DCAPE 4
3, Maintain a record of the Content Information and the Information Properties that it will preserve.	B1.1.2	Link submission and policy templates to the preserved collection	
4. Specify Submission Information Package format (SIP)	B1.3	Define templates that specify structure of a SIP and required content of a SIP.	DCAPE 3
5. Verify the depositor of all materials.	B1.4	Ingest data through a staging area that has a separate account for each depositor.	DCAPE 1
6. Verify each SIP for completeness and correctness	B1.5	Compare content of each SIP against template.	DCAPE 6
7. Maintain the chain of custody during preservation.	B1.6	Manage audit trails that document the identity of the archivist initiating the task	DCAPE 8
8. Document the ingestion process and report to the producer	B1.7	Send e-mail message to producer when process flags are set.	DCAPE 22
9. Document administration processes that are relevant to content acquisition.	B1.8	Maintain list of rules that govern management of the archives	DCAPE 10
10. Specify Archival Information Package format (AIP)	B2.1 B2.1.1	Define templates that specify structure of an AIP and required content of an AIP.	DCAPE 13
11. Label the types of AIPs.	B2.1.2	Store AIP type with each collection.	
12. Specify how AIPs are constructed from SIPs.	B2.2	Define transformation rule based on parsing of SIP template and AIP template	DCAPE 13
13. Document the final disposition of all SIPs	B2.3 B2.3.1	Maintain an audit trail for all SIPs	DCAPE 14
14. Generate persistent, unique identifiers for all AIPs.	B2.4 B2.4.1 B2.4.1.1 B2.4.1.2 B2.4.1.3	Define unique persistent logical name for each AIP	
15. Verify uniqueness of identifiers.	B2.4.1.4 B2.4.1.5	Identifier uniqueness enforced by algorithm that assigns identifiers	
16. Manage mapping from unique identifier to physical storage location.	B2.4.2	Storage location mapping enforced by iRODS data grid framework	

17. Provide authoritative representation information for all digital objects.	B2.5	Define template specifying required representation information.	
18. Identify the file type of all submitted Data Objects.	B2.5 B2.5.1	Apply type identification routine to each object on ingestion.	DCAPE 7
19. Document processes for acquiring preservation description information (PDI)	B2.6 B2.6.1	Define rule set that will be applied to extract PDI.	
20. Execute the documented processes for acquiring PDI.	B2.6.2	Apply PDI rules specific to a collection.	
21. Ensure link between the PDI and relevant Content Information.	B2.6.3 B2.7 B2.7.1 B2.7.2 B2.7.3	Set PDI extraction flag as part of PDI extraction rules.	
22. Verify completeness and correctness of each AIP.	B2.8	Compare AIP against template for required content.	DCAPE 14
23. Verify the integrity of the repository collections/content.	B2.9	Periodically evaluate checksums and compare with original checksum value.	DCAPE 17
24. Record actions and administration processes that are relevant to AIP creation.	B2.10 B3.1 B3.2	Maintain an audit trail of processing steps applied during AIP creation.	DCAPE 21
25. Specify storage of AIPs down to the bit level.	B4.1	Identify form of container used to implement an AIP.	
26. Preserve the Content Information of AIPs.	B4.1.1	Manage replicas of each AIP	
27. Actively monitor the integrity of AIPs.	B4.1.2	Periodically evaluate checksums.	
28. Record actions and administration processes that are relevant to AIP storage.	B4.2 B4.2.1	Maintain an audit trail of processing steps applied during AIP storage.	DCAPE 21
29. Prove compliance of operations on AIPs to submission agreement.	B4.2.2	Parse audit trails to show all operations comply with submission rule template	DCAPE 18
30. Specify minimum descriptive information requirements to enable discovery.	B5.1	Define submission template for required descriptive metadata.	
31. Generate minimum descriptive metadata and associate with the AIP.	B5.2	Apply rule to extract metadata specified within submission agreement.	DCAPE 11
32. Maintain link between each AIP and its descriptive information.	B5.3 B5.3.1	Package descriptive metadata within the AIP as an XML file	
33. Enforce access policies.	B6.1	Authenticate all users, authorize all operations	DCAPE 9
34. Log and review all access failures and anomalies.	B6.1.1	Periodically parse audit trails and summarize access failures	DCAPE 23
35. Disseminate authentic copies of records	B6.2	Define template to specify creation of a Dissemination Information Package (DIP)	DCAPE 26
36. Maintain replicas of all records, both content and representation information	C1.1.2	Periodically snapshot metadata catalog, and maintain at least two replicas	DCAPE 15
37. Detect bit corruption or loss.	C1.1.3	Periodically validate checksums	DCAPE 12

38. Report all incidents of data corruption or loss and repair/replace lost data	C1.1.3.1	Periodically synchronize replicas, and generate and store report	DCAPE 16
39. Manage migration to new hardware and media	C1.1.5	Replicate AIPs to new storage system	DCAPE 19
40. Document processes that enforce management policies	C1.1.6	Maintain copy of the rule base and micro-services used for each collection	
41. Document changes to policies and processes	C1.1.6.1	Version policies and micro-services	
42. Test and evaluate the effect of changes to the repository's critical processes.	C1.1.6.1.1 C1.2	Version state information attributes.	
43. Synchronize replicas	C1.2.1	Periodically synchronize replicas	
44. Delineate roles, responsibilities, and authorization for archivist initiated changes	C2.3	Define archivist roles and limit execution of preservation procedures to the archivist role	
45. Maintain an off-site backup of all preserved information	C2.4 B2.5.2	Federate two independent iRODS data grids and replicate digital holdings	
46. Maintain access to the requisite Representation Information.	B2.5.3	Manage Representation Information as metadata attributes on each record	
47. Maintain and correct problem reports about errors in data or responses from users.	B6.2.1 C1.1.1 C1.1.1.1 C1.1.1.2 C1.1.1.3 C1.1.1.4 C1.1.1.5 C1.1.1.6	Parse audit trails for unsuccessful operations and design appropriate micro-service recovery mechanisms	
48. Provide a search interface.		Provide a search interface.	DCAPE 24
49. Perform a virus check.		Perform a virus check.	DCAPE 5
50. Implement a loading dock.		Implement a loading dock.	DCAPE 2
51. Migrate records to new formats.		Migrate records to new formats.	DCAPE 20
52. Create and certify Dissemination Information Packages.		Create and certify Dissemination Information Packages.	DCAPE 25