

Assessment Plan - Preservation Worthiness

Given that the role of Fedora is to support an external preservation system, the main consideration is accessing repository content in formats suitable for preservation. There are five different methods for bulk copying repository content, which could be used to copy repository content to a preservation system:

1. backup/restore
2. export/import
3. retrieving content from the REST API (walking the repository)
4. retrieving content from the REST API (event-driven)
5. copying files from a filesystem federation

These mechanisms use a variety of different formats (MODE/ISPN/LevelDB binary, JCR XML, RDF, JSON) and use a variety of different workflows (externally-triggered, event-driven, automatic).

Questions:

1. What is the impact of using the different options on a running repository?
 - a. How do each of the methods scale?
 - b. What performance impact do they have?
 - c. What additional disk space, memory, etc. resources are needed?
2. How suitable for preservation are the different formats?
 - a. Can the datastream contents be accessed as/exported to files on disk?
 - b. Can the metadata be accessed in/exported to a human- and machine-readable format?

Testing Plan:

1. Ingest UCSD DAMS repository content (50K objects, ~8 TB) into Fedora 4.
2. Use each bulk-copying approach while running a performance test suite.
3. Examine output files from each approach to assess preservation value.