Fedora 3 to 6 Migration Community Updates

- NLM
  - Observations
  - Issues
  - Migration Tests
    - Repository environment #1: "collections".
    - Repository environment #2: "citations".
- Brown Univ
- University of Wisconsin - Madison
  - Observations
  - Issues
  - Migration Tests
    - UW Digital Collections Center Production Repository

NLM

Observations

1. External datastreams. Most of our binaries are of type E external. The migration tool migrates the Fedora objects, but not the type E external binaries (as expected). Thus we are left with object structure, metadata and RDF in OCFL format, but not the actual binaries themselves. If, how, when and where to migrate external binaries to an OCFL structure is TBD, but a major consideration for us in adopting OCFL.
2. Speed. The tool migrates objects at the rate of 15K-40K objects per hour. This should be manageable for our purpose.
3. For the "citations" repository, it consistently takes 30 minutes to build the datastream index before starting the migration. This server has 3.8M managed datastreams (1 per object). The option to cache this index when resuming migrations is helpful.
4. CPU time. Consumes about 30%.
5. Layout. In flat and pairtree migrations the PID is used to form the path; for example PID nlm:nlmuid-101588995-bk (stored FOXML file name nlm_nlmuid-101588995-bk) becomes /ocfl/nl/m+/nl/mu/id/-1/01/58/89/95/-b/k/5-bk. Characters such as – are problematic in Linux. See **FCREPO-3180** - Getting issue details... STATUS .
6. It would be nice to declare use of another field, or input map, to dictate the value to use for layout path generation. For example, it may be nice to use 101588995_bk to generate a path for PID nlm:nlmuid-101588995-bk. Also included in **FCREPO-3180** - Getting issue details... STATUS .
7. Migrated datastreams have no file extension. It would be nice if migrated datastreams have a file extension inferred from the MIME type; e.g. DC.xml instead of just DC, and OCR.txt instead of just OCR. This should particularly help out with in-line XML datastreams. **FCREPO-3181** - Getting issue details... STATUS
8. OCFL versions appear to be created based on datastream timestamps. Each unique timestamp creates a new OCFL version, even if they were part of the same Fedora version in the AUDIT trail and differed only by milliseconds.
9. Add XML declarations for migrated in-line datastreams. **FCREPO-3197** - Getting issue details... STATUS

Issues

1. Occasional "Unable to delete staging" messages, which stops a migration. See **FCREPO-3187** - Getting issue details... STATUS and **FCREPO-3191** - Getting issue details... STATUS .
2. Migrated objects from the "collections" repository generate approx. 40 files in OCFL from a single Fedora 3 FOXML file. This repo with 4M objects could generate 160M files. We have already run out of filesystem inodes when attempting migration runs. It may be useful to optimize the number of generated files to mitigate this issue.
3. Migrated external datastreams are not resolvable. **FCREPO-3198** - Getting issue details... STATUS

Migration Tests

This section logs migration tests for our two different repository environments, "collections" and "citations".

Repository environment #1: "collections".

Approx. 4.3M records in legacy format, mostly books, pages and still images. Objects have many datastreams, binaries are generally type E external.

Fedora 3.8.1. VM with 4 cores, 8 GB RAM. CentOS release 6.10 (Final), Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz

Number of objects	Execution Time	Source Layout	Dest. Layout	Migration tool version	Notes

1000	4 min	legacy	pairtree	11/26/19	1K fedora items produced 42K+ files
1000	3 min	legacy	truncated	11/26/19	1K fedora items produced 43K+ files
100,000	6.5 hours	legacy	flat	11/26/19	
1 million	~3 days	legacy	pairtree	11/26/19	Execution crashed twice for "unable to delete staging" file issues, resume option had no issues running
full run (4,656,669 items)	7 days	legacy	pairtree	2/4/20	No issues observed for successful full migration run. Required deployment of new filesystem with large inode limit.

Repository environment #2: "citations".

Approx. 3.8M records in akubra format, all citations with one small type M XML datastream (the citation payload).

Fedora 3.8.1. VM with 1 core, 8 GB RAM. CentOS release 6.10 (Final), Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz

Number of objects	Execution Time	Source Layout	Dest. Layout	Migration tool version	Notes
2000	32 min	akubra	pairtree	11/26/19	Includes 30 min to build the index. Hung on completion-could not delete index.
10,000	42 min	akubra	flat	11/26/19	Includes 30 min to build the index. Hung on completion-could not delete index.
554,695	13 hours	akubra	flat	11/26/19	Attempted to migrate 1M records. Includes 30 min to build the index. Crashed due to UnrecognizedPropertyException.
full run (3,830,777 items)	5 days	akubra	truncated	2/4/20	No issues observed for successful full migration run.

Brown Univ

University of Wisconsin - Madison

Observations

1. Storage environment: for the purposes of this test (and for our real migration), we are migrating from one CIFS-mounted remote filesystem to another CIFS-mounted remote filesystem.
2. Datastream index: takes about 1h10m minutes to build, and occupies 327MB of disk space.
3. Source layout. Akubra hash storage, using the pattern "#/##/##" for both datastreams and objects.

Issues

Migration Tests

UW Digital Collections Center Production Repository

Fedora 3: Approx. **561,000 objects (559GB)**: mostly books, pages and still images, with some audio, video, and PDF resources. Approximately **2.36 million datastreams (10.3TB)**. Content objects have one binary datastream and 5 XML metadata datatstreams. Container objects have ~5 XML metadata datastreams. All datastreams are either inline or managed (no external or redirect datastreams).

Fedora 3.8.1. Migration run on VM with 4 cores, 8 GB RAM. CentOS Linux release 8.2.2004 (Core), Intel(R) Xeon(R) Gold 5220 CPU @2.20GHz

Command run:

UW Madison migration-util command line

```
$ java -jar target/migration-utils-4.4.1-SNAPSHOT-driver.jar --migration-type=FEDORA_OCFL --source-type=akubra
--datastreams-dir=/fedora3-prod/fedora/datastreams --objects-dir=/fedora3-prod/fedora/objects --target-dir=
/fedora-migration-test --index-dir=/var/tmp/datastream-index
```

Number of objects	Execution Time	Average seconds per object	OCFL repository size	Source Layout	Migration tool version	Notes
1000	Datastream index: 1h17m OCFL repo: 4h36m	16.3 sec	184GB	Akubra	07 Oct 2020 (81586bf)	with param --pid-file=1000pids.txt datastream index cleared after run
10,000	Datastream index: 1h5m OCFL repo: 11h48m	4.3 sec	688GB	Akubra	07 Oct 2020 (81586bf)	with param --pid-file=10000pids.txt datastream index cleared after run
100,000	Datastream index: 1h9m OCFL repo: 3d20h16m	3.3 sec	6.8TB	Akubra	13 Oct 2020 (4a9f19c)	with param --pid-file=100000pids.txt datastream index cleared after run
All 561,000	Datastream index: 1h10m OCFL repo: 20d21h12m	3.2 sec	39TB	Akubra	23 Oct 2020 (43b7bae)	all pids