2022-08-18 - Fedora Tech Meeting

Time/Place

This meeting is a hybrid teleconference and slack chat. Anyone is welcome to join...here's the info:

- Time: 11:00am Eastern Daylight Time US (UTC-4)
- Audio/Video Conference Link:

  https://lyrasis.zoom.us/j/81195576163?pwd=MkIvektGQVhYdGpkZndNOE1DbG5UZz09

  - Meeting ID: 811 9557 6163
  - Find your local number: https://lyrasis.zoom.us/u/kKttE88vD
  - Join us on the #Tech channel on Fedora Slack: https://forms.gle/MjgCvNch7eaRoXRb8

Attendees

1. Danny Bernstein
2. Arran Griffith
3. Jared Whiklo
4. Ben Pennell
5. Michael Ritter - Meeting Chair
6. Demian Katz
7. Jon Roby ⭐
8. James Alexander
9. Calvin Xu

Each week a meeting chair will be assigned based on a rotating schedule.

⭐ - denotes note taker

Agenda

1. Announcements
   a. Recap from last week on User Information Gathering

1. New tickets:

FCREPO-3837	Improve Feedback in Validation Tool

2. Updates on Backlog Tickets:
3. In review tickets:

FCREPO-3836	Migration-utils generate invalid RDF triple if xml:lang is present

FCREPO-3835	Unused versioning actions in web ui

FCREPO-3833	Update Head Only Validation

4. Other topics:
5. Discuss migration:
   a. Demian's Migration

Notes:

Announcements
  Recap from last week on User Information Gathering
  Danny Bernstein - everyone thought #4 was a no-brainer, should do that anyway.
     "phone home" wouldnt be useful, reward would be low, reputational risk would be high.
     stats export feature - able to voluntarily share with community
     call to action (banner) everywhere -
  Arran Griffith - where would that information go?
      - endpoint to receive info on fedora/lyrasis?
  Arran Griffith - lyrasis is reconfiguring current registry
     will take these to governance as options
  Danny Bernstein - return header and call to action would be easiest option
  Danny Bernstein - will create tickets, 1 for UI work, 1 for API work, 1 that might be broken up for stats collection
  Arran Griffith - Sept 8th meeting, Danny Bernstein will connect and explain options/reason

Arran Griffith - fedora developer position is live, live until 29th, lots of interest

  New tickets:
    FCREPO-3837
    Improve Feedback in Validation Tool
  Michael Ritter - improve logging during report generation, should be complete soon
  Demian Katz - still an issue with escaping double quotes?
  Michael Ritter - will have to look at it again
  Demian Katz - migrated data looks ok, but extra slash in validation tool?
  Michael Ritter - should be fixed in current validation tool
  Demian Katz - tool version used is nov 11 2021
                          - CI is failing on project, something to look at
                          - thanks for the tool, helps a lot with confidence in migration

  Updates on Backlog Tickets:
    In review tickets:

FCREPO-3836
  Migration-utils generate invalid RDF triple if xml:lang is present
Jared Whiklo - was not packaging jena properly, was using assembly plugin, replaced with shade plugin and calls
    correct n-triples writer as it was using 2 writers and 1 is broken but unused
    added new PR, no new code, just changes packaging.

FCREPO-3835
  Unused versioning actions in web ui
  closed - Danny Bernstein sabotaged Arran Griffith by merging this ticket and closing it

Danny Bernstein - added commit from FCREPO-1994 in jira
Jared Whiklo - was found during RC but not important enough to be added to RC? few simple fixes

FCREPO-3833
Update Head Only Validation
closed - Danny Bernstein merged and closed ticket
Michael Ritter - has extra validation, checking num of objs in fedora 3 vs fedora 6 when processed.
    head only validation with f3 doesn't count deleted objs, so count might be larger on f6 than on f3
    might have PR for this in a few days

Other topics:
  Discuss migration:

Demian's Migration
Demian Katz - test env is setup correctly, migrated & reindexed 600,000 objs, took ~12 days. only 2 problems.

1) fedora exception during reindexing, happened 32 times over 3 days while reindexng 600k objs.
occurs in pairs, 5 workers reindexing simultaneously performing only reads, wildcard exception mapper exception,
did reindex on that object and it worked afterwards. maybe resource utilization?.

2) no luck with camel toolbox reindex, maybe lack of understanding in activeMQ, wrote script to do reindex. tried camel, seemed to do 20k objs then
tailspin.

Jared Whiklo - spent too much time working on forwarding AMQ endpoint to AMQ endpoint as UofM use it. console for AMQ to look under the hood
did Demian Katz change the topic to a queue? demian did not

topic is non-durable state, works of pub/subscribe model, queue is durable (persistant msgs on disk?)
topics should be wiped clean between restarts, sounds like demian ran out of space?
(web) consoles are good to check on status, msgs being processed
demian concentrating on pids, 600k pids not datastreams, reindexing shouldnt care about datastream reindex messages
Jared Whiklo - able to filter out binaries? trivial to add in filter for only messages demian needs (top level PIDs, not datastream, not binaries)
Danny Bernstein - what is median processing time per pid? demian - doesnt matter as its queue related
Demian Katz - queue is being added to, tool looks through queue for dupes, so time increases as both increase
Jared Whiklo - filter exists to filter by container
    switch to queue to make it persistent between restarts?
    setup external AMQ as monitoring tool for reindex service msgs?
Demian Katz - can the reindexer be throttled? Jared Whiklo - items added as fast as it can, throttling is on receiver end
    next step, switch from topic to queue and see if that works
    change to external AMQ to catch msgs/errors/queue status
Demian Katz - will keep the group posted on experiments
    tried to setup standalone AMQ but it didnt work
Jared Whiklo - maybe from using path and not file:// in configuration
Demian Katz - why is there an exception from using triples in XML format
Jared Whiklo - exception maybe from iso-8859 encoder, maybe big string/complex language,seems to be stateful exception
Demian Katz - nothing too crazy in objects, so shouldnt be causing an exception
    reindexer pulls triples in XML format, so that could be the reason?
Danny Bernstein - not the first time we've seen this issue, create a ticket for this issue
Demian Katz - will do some research for more info and create ticket