2019-07-23 Docuteam Fedora 6 Pilot Meeting

Time/Place

Time: 12:00pm ET

Audio Conference:

Join from PC, Mac, Linux, iOS or Android: https://duraspace.zoom.us/j/8128353771

Or iPhone one-tap :
US: +16699006833,,8128353771# or +16468769923,,8128353771#
Or Telephone:
Dial(for higher quality, dial a number based on your current location):
US: +1 669 900 6833 or +1 646 876 9923
Canada: +1 647 558 0588
Australia: +61 (0) 2 8015 2088
United Kingdom: +44 (0) 20 3695 0088
Meeting ID: 812 835 3771
International numbers available: https://zoom.us/u/MO73B

Attendees

- David Wilcox
- Andrew Woods
- Danny Bernstein
- Thomas Bernhart

Agenda/ Notes

Topic	Lead
Review pilot guidelines and answer any questions	David
Discuss Docuteam pilot details - What are the desired outcomes? - What are the timelines? - What is our communication plan?	David
Next steps	All

Notes

- Review of Pilot objectives
- Desired outcomes?
    - Data model would change when migrating to Fedora 6 in all probability
    - Will have specific needs for the data migration
    - DT is considering moving to a linked data
        - there is a proposed mapping - Thomas will need to dig it up
    - Clear picture of how they will migrate to Fedora 6
    - Clear picture of what it will take in terms of time and resources to migrate
    - Clear picture of how to deploy Fedora 6
    - Specific performance requirements: not clear what will be necessary
    - Size: 100 GB binary - 5 TBs of binaries
    - What is object count range across instances? *
    - Ingest and retrieval speeds? Data / Object count per second
    - 50 fedora installations
    - Objectives related to surrounding processes or front-ends:
        - will be using some kind of index
        - will be using a triplestore
        - currently using CRUD functionality
        - what are the queries you expect to be supported? *
- Timelines:
    - Starting in September
- Our expectation is that we'll be able to get feedback on alpha / beta releases
- Communication

- We will provide regular (weekly? bi-weekly) check-ins for pilot partners
- meeting at 11 am eastern would be ideal
- Slack channel: #fedora6-pilots
- Is Docuteam available for development work?
  - Testing yes - unquestionably
  - Development: we don't know - looking at outsourcing development.
  - May be able to contribute to migration libraries.
- Providing sample data to the Lyrasis/Fedora team ASAP would be appreciated.
  - DT has about 20 sample objects that are based on their objects that they can provide.
    - DT: look for it in the second week of August.

Use cases and requirements

General requirements

- we must be able to run Fedora 6 on Windows Servers

Query service

Use cases

- if a query service is integrated into fedora, docuteam would like to use this service for the following queries:
  - total number of objects in a namespace (used to be PID namespace in Fedora3; equivalent to PID namespace in Fedora6 has to be defined; possible solutions: a) use toplevel objects to group objects by namespaces, b) use a rdf triple to assign a object to a pid namespace, c) use the Name Assigning Authority Number (NAAN) of ARKs to group objects)
  - get all available file formats in a fedora instance (we use PRONOM identifiers PUIDs to store this information) and to number of objects with a specific file format
  - get the total space used
  - get available free space for the persitent storage (optional as this could be solved by monitoring tools)
  - get the total space used by namespace (see above on what a namespace could be)
  - get all objects with a specific Triple (e.g. PRONOM identifier, some alternative identifier)

Requirements

- use existing standards for the query language (SPARQL and LDPath (https://marmotta.apache.org/ldpath/) come to mind)
- should support aggregation functions (e.g. sum(), count())
- should reuse existing indexes like Solr or triple store if present?
  do not build another index, if these tools are present anyway

Questions/Comments

Currently it's rather unclear for us, how this query service would be integrated into fedora and what are the use cases that such a service would support. We are skeptic that a query service would support all our use cases and that it adds an additional index for information, that would already be present in a triple store and/or Solr index. Therefore we think such a query service should be either an optional component that can be installed along Fedora6 (similiar to a fixity service) or be able to use existing technologies like a triple store or Solr/Lucene index as a backend.

However we would see a use case, to extend the existing simple web UI of Fedora5 with a possibility to search for objects (similiar to the search interface that what was available in fedora3 when accessing the endpoint /fedora/objects). As mentioned above, such a UI could leverage existing indexes.

- make optional?
- will query service be part of the fedora API spec?
- does it make sense to hardcode the supported queries?
- only provide a search UI but actually use a triple store?

Answers/Remarks by **Andrew Woods**

- Regarding "Query service", we have gotten consistent feedback that an integrated, synchronous search index should come with Fedora. The use case is for clients to create/update a Fedora resource, then immediately be able to query Fedora with the expectation that the resource be in the index. The externalized, asynchronous indices do not satisfy this use case.
- We will certainly use the queries you have enumerated in the testing of the new query service.

Persistence

Use cases

- For our cloud infrastructure, docuteam would like to use s3 compatible object storage (provided by Ceph https://ceph.io/)
- Some docuteam clients might want to stick with a simpler storage model than OCFL to reduce storage requirements

Requirements

- OCFL backend: native support to use s3 compatible object storage
- alternative "simple" persistence implementation (optional?)

Migration

Use cases

- Wtih the switch to Fedora6, docuteam wants to switch to a ontology-based data model.
- Docuteam would like to leverage the functionality of a generic migration utility.
- Docuteam would like to migrate Fedora 3 XML-Datastreams into triples.

Requirements

- possibility to select/configure which Fedora3 datastreams should be converted to Fedora6 binaries
- possibility to extent migration utilities with custom parsers/functionality to create triples

Documentation

Requirements

- Step by step installation manual for production use
- Recommendations for storage systems (e.g. WORM, Object Storage, generic NAS)

Communication channels:

- fedora-project.slack.com:
- wiki
- #fedora6-pilots
- #fedora-tech
- fedora-tech@googlegroups.com
- fedora-community@googlegroups.com

Actions

- Thomas Bernhart to deliver sample data to Andrew Woods by August 16th.
- Thomas Bernhart to deliver use cases, and requirements to Andrew Woods by September 6th.