Spoken Word internal wiki

Workflow Notes prepared for Fedora UK&I Meeting December 2010

Notes prepared for Fedora UK&I Meeting December 2010

Last modified by Iain Wallace on 2010/12/10 12:39

Migration to Fedora

Spoken Word media has been served from a legacy PHP/MySQL repository for several years. This content has been successfully migrated to Fedora. The legacy system is currently still used for ingest and discovery but this will move to Fedora/Hydra (see below). Specific developments:-

- Data cleansing and rationalisation
- Compound content models (all available at: https://github.com/SpokenWordServices/Miscellaneous-Fedora-stuff/tree/master /content_model_objects)
  - gcu:swvideo
  - gcu:genericvideo
  - gcu:swaudio
  - gcu:genericvideo
- Descriptive metadata in MODS
  - Dynamic transform to oai_dc for OAI-PMH with Proai
    - Currently being harvested for GCU instance of SerialsSolutions Summon and Mediahub
- Fedora 3.4 upgrade
  - Upgrade required to get round FCREPO-704 (upload of files larger than 2GB fails via REST interface)
  - Also a chance renew infrastructure and standardise on Tomcat 6 and Debian 6.
  - Managed content storage now on an EMC SAN (being served via NFS)
    - fedora-rebuild.sh bug: classpath issue on Debian 6 meaning rebuild script fails. Also experienced by York. Under investigation.
- Outstanding issues with Fedora development:
  - Lack of certain HTTP headers (Content-Range etc.) still prevents media being served efficiently - this is scheduled for 3.5

Metadata issues

The default set of Fedora object relationships does not provide sufficient richness to express the different types of semantic refinements specific to broadcast/timed media. We want to capture this richer relationship information and make it searchable via the Resource Index. These new relationships are intended to be used in tandem with the existing Fedora schema.

- Draft new relationship ontology for timed media:
  - https://github.com/SpokenWordServices/Miscellaneous-Fedora-stuff/blob/master/ontologies/sw-object-relationships-schema.xml
  - Derived from BBC Programmes Ontology:
    - http://www.bbc.co.uk/ontologies/programmes/
  - Request for comment, particularly:
    - Is this schema more generally applicable to all timed media?
    - Does the terminology make sense outside broadcast media?
    - Can it be used for non-timed media, e.g. images or text?

Hydra

- Overall impressions so far
  - Rapid development; productive community atmosphere
  - GitHub as a development venue a major boon to collaboration
  - Rails, ActiveFedora, OM and Blacklight stack a pleasure to work with
  - Stable and fast so far (1ms query times = good)
  - But needs much more testing in practical environments (i.e. not just unit tests)
- Completed work
  - Conversion of various Hydra models from DC metadata to MODS: GenericImage, GenericContent etc.
  - Draft new audio model, GenericAudio (to be followed by GenericVideo)
  - Uses HTML5 playback with fallback to Flash player
- Ongoing work
  - Revision of the way Hydra stores information about uploaded files: separation of this information into a distinct datastream will give the flexibility to merge compound and atomic aspects in a single object (e.g. thumbnail inline, link to external media file).
  - Content models to support (a)synchronous generation of derivatives from uploaded master file via invokation of external web services (using derive_all method call)
    - Might also be a useful approach for preservation metadata?
- Demo URLs:
  - Example of an image with MODS metadata at:
    - http://catalogue.spokenword.ac.uk/catalog/changeme:41

- audio at:
  - http://catalogue.spokenword.ac.uk/catalog/changeme:40
- content at:
  - http://catalogue.spokenword.ac.uk/catalog/changeme:43

Work related to a possible timed media Hydra 'head':

- HTML5 Timed transcript alignment tool
  - Plays synced W3C Timed Text to media
  - Automatic transcript time alignment via YouTube machine learning sync tool: conversion scripts to generate W3CTT.
  - Demo: http://resources.spokenword.ac.uk/testing/captions/dfxp-example.html
  - Next steps: codify into Hydra model and add necessary view templates
- Real-time video thumbnailing
  - Rudimentary web service using ffmpeg to generate video thumbnails - works like ImageManip.
  - Code at : https://github.com/SpokenWordServices/Miscellaneous-Fedora-stuff/tree/master/video_thumbnailer_service

Other development work:

- Missing files checker - useful for verifying migrations to Fedora. Compares two file structures and compares every file by checksum, reporting on any which are present in the 'source' but not the 'destination'. Tolerant of file renaming.
  - Code at https://github.com/capncodewash/Misc-shell-scripts/blob/master/find_missing_files.sh

Further reading:

- Spoken Word Fedora repository: http://repo.spokenword.ac.uk/fedora/search
- Spoken Word GitHub repositories:
  - https://github.com/SpokenWordServices/Miscellaneous-Fedora-stuff
    - Contains content model objects, xacml policies, ontologies, collection objects and OAI system objects
  - https://github.com/SpokenWordServices/hydrangea
    - Contains our fork of Hydrangea with various customisations and improvements, e.g. GenericAudio

Tags:

Created by Graeme West on 2010/12/10 11:47

No comments for this document