OR2018 24x7 presentation Submission

13th International Open Repositories Conference, June 4th-7th, Bozeman, Montana, USA

Digital Preservation through EPrints-Archivematica Integration

Tomasz Neugebauer, Concordia University, tomasz.neugebauer@concordia.ca; Justin Simpson, Artefactual Systems Inc., jsimpson@artefactual.com; Justin Bradley, University of Southampton, jb4@ecs.soton.ac.uk

Session Type

- 24x7

Abstract

This presentation addresses digital preservation challenges with EPrints repository content through integration with the Archivematica system specifically designed for digital preservation. A workflow and folder structure using BagIt for exporting EPrints content into Archivematica is described. A sample item export with multiple files and formats is used to demonstrate the integration plan.

Conference Themes

- Open source software - sustainability of software developed locally and large open source systems, legacy code
- Content - research data, digital preservation, persistent urls, archiving
- Infrastructure/Integrations - integrations between systems, changing technical environments
- Challenges of sustainability - funding, local, technical, community

Keywords

Digital Preservation, Integration, EPrints, Archivematica

Audience

The presentation is of interest to repository managers, developers and librarians looking to deliver digital preservation functionality with EPrints, as well as other repository systems. While various methods for DSpace, Islandora and Dataverse integration with Archivematica already exist, the proposal is new for EPrints and an opportunity to compare the digital preservation integrations across different open access systems.

Background

The conference theme is focused on the question of “how, why, and what it will take to make open sustainable?” and this work proposes the answer that sustainability is achieved through the integration of specialized open source system. Digital preservation, defined as the methods to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term, requires sustainability. EPrints prioritizes open access, but the current state of digital preservation functionality in that system is insufficient and unmaintained, while Archivematica specializes in precisely this functionality.

Presentation content

Improving the digital preservation workflows for the EPrints Spectrum research repository (https://spectrum.library.concordia.ca/) is one of the objectives of the Concordia University Digital Preservation Working Group. Most of the digital preservation functionality available to the EPrints community is the result of the 2005-2009 Preserv project (http://preserv.eprints.org/). The Digital Preservation Toolkit (http://bazaar.eprints.org/142/) that was released in 2011, is out-of-date and incompatible with the latest version of dependencies (e.g. DROID). Meanwhile, since 2009, Archivematica has emerged as the leading open source solution specifically intended for digital preservation with an active community that has been able to maintain and develop the standards-based system.

Full digital preservation support requires file format verification, validation, migration/normalization for preservation purposes, along with an extensive METS file containing standardized (e.g., PREMIS) preservation metadata. The sustainable solution for EPrints is to leverage the digital preservation Archivematica software to deliver this functionality, rather than attempt to duplicate the effort within EPrints. The proposed integration is to export out of EPrints all live archive eprints and package them along with the digital files using BagIt. Archivematica imports these bags containing metadata and data, and processes them to generate AIPs (Archival Information Packages) with their corresponding METS files.

The following is a summary of the proposed workflow for EPrints-Archivematica integration:

- "Digital Preservation Export" batch script runs periodically that identifies new/updated items to export and generates the bags described below.
- BagIt hierarchical packaging format is used create one bag per eprint, including: a **metadata folder** with Dublin Core metadata (in TEXT format), EPrints XML metadata, all uploaded digital files that are a part of the eprint in a **data folder**, all EPrints generated "revision" XML files in a **revisions folder**, and a **derivatives folder** containing any derivative access files that were generated by EPrints, such as thumbnail images, audio access files, video access files. Bags are moved to specified shared storage location.
- The following would be the structure of the **data folder**:

eprintid-XXXXX → documents → documentid-XXXXX → files → fileid-XXXXX → folder# → filename

- The following would be the structure of the **derivatives folder**:

fileid-XXXXX → folder# → filename

- Archivematica monitors shared storage for new bags and processes new items by: extracting the Dublin Core Metadata (in TEXT format) and converting it to JSON for indexing, using the eprintID as an indexed identifier for the AIP, generating digital preservation AIP by running the full identification, characterization, normalization routines on all files in the bag, and finally moving the AIP to archival storage.

This integration is currently in the technical specification phase without an implementation, but the presentation will include a sample export with multiple files in MIDI, WAV, PDF, PPTX and TXT format.

Repository System

- EPrints
- Archivematica

Conclusion

EPrints is a state-of-the-art open access repository system, while Archivematica is a web and standards based solution for digital preservation, with a sustainable open source model. This presentation describes a new integration proposal using a custom content and metadata export from EPrints to Archivematica. Once implemented, Archivematica would create AIPs for all new and updated eprints in the repository.