

Development & Deployment Team

- Katherine Lynch (UX) Senior Application Developer
- Michael Gibney (Git & git-annex) Senior Application Developer
- Martin Oestergaard (Ceph) Unix Systems Administrator

Repository Ecosystem



Colenda is intended for use as part of a larger repository ecosystem. The team at Penn Libraries is currently in the process of creating and test-deploying an architecture using a Ceph storage cluster, git-annex for versioning of large files, and Hydra/Rails for management of object files and metadata representation.

- **Colenda's UI/UX**
Colenda uses Hydra for metadata aggregation in Fedora, as well as search and discovery in Solr/Blacklight. Additional functionality to curate and manage assets and metadata is provided in the application through Rails. Objects' descriptive metadata is ingested to Fedora, and structural metadata referenced for display via the Rails application.
- **Git & git-annex**
Git repositories created through Colenda's workflows use git-annex, a tool that extends Git to version files for long-term preservation without checking the binaries into Git; rather, git-annex versions references to file content and location. Binaries are stored in a key/value store where the key is derived from a SHA-256 checksum of file content. This provides a layer of abstraction that allows users to flexibly rearrange items on the file system without breaking ingestion workflows. Files are stored on separate storage identified as a git-annex special remote. A special remote is an abstraction that handles storage of binary content but not versioning metadata, and integrates with the Git ecosystem.
- **Ceph Storage Cluster (S3)**
The binaries referenced in Colenda's Git repositories live on remote storage that runs as a Ceph storage cluster. Ceph is a hardware-agnostic software storage platform that prioritizes data replication across multiple nodes for high availability, high replication for fault-tolerance, and scalability. The versions of the files targeted for long-term preservation are stored in one place and referenced using git-annex and Fedora plugins on the remote, using an S3-compatible gateway API.

Penn Libraries are creating a long-term preservation ecosystem including Colenda at its core. Colenda is a metadata-agnostic application that provides materials-agnostic functionality for distributed workflows around administration of digital assets and metadata through a pluggable architecture for metadata schemata and entry. Objects consist of two types of things: assets and metadata. Assets consist of all files associated with an object's representation. Metadata consists of information describing the object or its structure. In Colenda, metadata sources come in the form of spreadsheets that directly contain metadata and/or information that allows metadata lookup services to extract bibliographic information for descriptive metadata from external sources such as the library catalog. Because the application accepts files as metadata sources, this allows the metadata entry interface to be versioned alongside the metadata it is creating, which allows an ongoing thorough understanding of the metadata based on the context in which it was entered.

The beginning scope of the project addresses page-turning objects, specifically manuscripts from Penn's collections, and represents significant development of a completely generalized approach to object creation and ongoing management. Its first phase of software development has just been completed and the team is currently working on crafting a robust stack of software and hardware fitting together with this application.

