**Medusa**

**Medusa Workflow and Preservation System**

The Scholarly Communication and Repository Services development group at the University of Illinois at Urbana-Champaign library is building a cradle to grave workflow management and preservation repository system based on the Hydra technical architecture. In keeping with Hydra’s mythological theme, we named our system Medusa. Medusa uses Fedora for its repository layer, Apache Solr for full text indexing, and Hydra tools — ActiveFedora, Solrizer, and OM — for creating and managing objects in the codebase. The workflow engine calls on a series of disparate micro-services via the Advanced Message Queuing Protocol (AMQP), while moving an information package along the process from digitization to archival storage, generating PREMIS metadata along the way. The Medusa workflow model is loosely based on the Archivematica project, but is written entirely in Ruby and is designed to be deployed in a distributed environment.

Even the simplest digital objects can be created in multiple stages, oftentimes by more than one person and/or department, before they are ready to be ingested into a repository or digital library management system. Before Medusa, workflows for creating digital content at the library would generally involve storing and accessing files on a shared file system, and could be standardized or ad hoc. A typical workflow might include scanning an image or page, cropping and de-skewing the image, creating screen-size and thumbnail variants, and metadata creation, all happening before the package undergoes any preservation activities, leaving the files open to risk of corruption while the object is being built. The Medusa approach aims to address this problem by beginning digital preservation at the very start of the process.

With Medusa, a Fedora digital object is created as soon as the image is scanned, the ETD is submitted, or the audio stream is recorded. All other work on the object happens on the Fedora object itself — metadata streams may be added to the object later, as well as cropped and de-skewed versions of the content or other derivatives. Even if the original content in the object is changed, the Fedora system performs date-stamped versioning of the content and can generate provenance statements automatically. Fixity checks and checksum generation can also be done automatically with Fedora. With Medusa, from the moment the digital object creation process has begun, so has digital preservation.

As mentioned above, the Medusa workflow process is made up of several micro-services. These services include bitstream validation, virus checking, technical metadata creation, and handle assignment, and each of these services creates its own provenance statements, which are stored in the object as a PREMIS XML stream. Modeled after the Archivematica workflow, the path of micro-services is forked near the end of the object-creation process, such that two versions of the object are created: a dissemination information package (DIP) and an archival information package (AIP). The DIP can be customized for ingest into a specific digital library software system — ContentDM, ArtStore’s SharedShelf, and DSpace, are examples. The AIP will remain in long-term archival storage in Fedora.

**Current Status:** In development. Pilot image collections are being migrated from ContentDM to Medusa now. We expect we will be ready to accept other content by fall 2011.