Using DSpace for Publishing Electronic Theses and Dissertations

James Halliday and Randall Floyd
jhallida@indiana.edu, rdfloyd@indiana.edu
Indiana University

The IUScholarWorks Repository is a DSpace-based institutional repository for the dissemination and preservation of Indiana University's scholarly output. Some time ago, our team made a decision to incorporate electronic theses and dissertations (ETD's) into our DSpace repository, and this created several technical challenges for us. Getting ETD's into DSpace is a challenge that a number of institutions have tackled recently, and several innovative solutions have been found, such as Vireo, the ETD submission management tool from the Texas Digital Library. However, we were faced with a number of requirements in our ETD workflow that had not yet been encountered by other institutions, and required some interesting solutions.

In this proposal, we will provide an outline for a presentation that will discuss these challenges, and the solutions that were envisioned. We will also provide an update on our current progress towards implementing our plans, and discuss the future work that is left to be done.

Background

Indiana University planned and implemented a pilot institutional repository called IUScholarWorks in early 2006. By mid- 2006 the service went live, which initially contained six active communities and 124 documents. As the repository grew from a small pilot project into a full-fledged repository service, the need became pressing to incorporate ETD's into the repository service. PhD candidates of the IU Graduate School already submit their completed dissertation documents to ProQuest for archiving and publishing, but there was a desire for a free and open in-house repository alternative, a platform for publishing that students could choose in addition to the usual ProQuest submission process. The IUScholarWorks DSpace repository provided a great solution for this need. However, we were faced with a number of challenges in implementing something that would work with our specific needs. First, we needed an easy way to automatically ingest the dissertation content, with little or no interaction with the submitter. We also needed the ability to embargo content (both content and metadata) until the appropriate permissions could be obtained. Lastly, we needed a way to integrate the IUScholarWorks dissertation object with the record for the dissertation in our library's online catalog (IUCAT). The following sections will examine each of these challenges in more detail.

The First Challenge: A Drop Box Service

As mentioned earlier, the Indiana University Graduate School dissertations are submitted to ProQuest upon completion. ProQuest provides additional formatting and other services, but also returns to us the original submitted documents. It is these original documents that we have the right to publish in IUScholarWorks, with only the original author's permission.

It became apparent that we needed to create a staging environment, a place where the raw incoming documents could be temporarily stored and later processed for ingestion into DSpace. This staging step is necessary for any automated batch ingestion of content into DSpace, whether through SWORD or the DSpace ItemImporter.

Our original conception was for a dissertation-specific drop box system, where we could take the dissertations received from ProQuest and automatically ingest them. However, it became immediately apparent that the automated part of the workflow for this one case could be generalized and that we should instead create a reusable system. The scope of this project was then expanded to create a service that could process and transform any kind of content in various kinds of workflows. The result was a new automated workflow system that extends our service called The IUScholarWorks Drop Box Service.

The IUScholarWorks Drop Box is a simple service that users see as directories on a server corresponding to collections in IUScholarWorks. Depositing items using the Drop Box is just a matter of uploading packages to the appropriate directories; these items will then automatically appear in the appropriate collections after automated processing takes place. It is the responsibility of the user to prepare these item packages, which consist of the actual item files and corresponding metadata in a predetermined XML format.

This service provides an alternative to online submission and can be used by anyone or any system that can fabricate XML metadata and upload files to a server. Item packages could be placed in the drop box as an event in an automated workflow or manually by an individual. In either case, it is anticipated that metadata for content might already be expressed in an XML format that is appropriate for that content, or that it is more natural for the user to create a new neutral format than to create metadata specific to DSpace. The Drop Box Service is then responsible for transforming the content-specific metadata into the DSpace Dublin Core format expected by the Item Importer. Handling the metadata this way also gives us the opportunity to properly analyze and address metadata concerns and usage before being confronted with loading into DSpace.

The actual system that facilitates this service is called the IUScholarWorks Drop Box Processor. Essentially, it is a pre-processor and control system written around the DSpace Item Importer. Its job is to discover the existence of content in drop box directories, to figure out which DSpace instances and collections content goes into, to transform content-specific metadata into DSpace DC, and to stage the content appropriately before finally launching the Item Importer against the prepared items.

The Drop Box Processor is configurable on a per-collection basis via XML configuration files, and allows the IUScholarWorks administrator to configure drop box directories to have different behaviors and outcomes. When configuring a directory, the administrator can specify which DSpace instance to import items into, which collection to import items into, which DSpace e-Person to import as, and which style sheet to be used to transform content-specific XML metadata into DSpace Dublin Core.

The Drop Box Service is currently in production with IUScholarWorks. It has been used to ingest dissertations into a private dark archive, and is currently being used to regularly ingest content into our largest live collection.

The Second Challenge: Adding Embargo Functionality

The second challenge involves our need to embargo certain content within the DSpace system. As mentioned above, our first test with the Drop Box Processor involved ingesting content into a completely non-public dark archive. In order to bring our ETD workflow into production, we needed a way to hide certain content from view, and to have our Drop Box Processor automatically ingest dissertations with an embargo. We also wanted to give dissertation authors the option to have their work embargoed for a certain period of time before making their work public in IUScholarWorks.

We decided that we wanted to make the embargoed content completely hidden from public view, meaning that the bitstreams should not be accessible, but also that the record metadata should not be accessible, either through direct record access, or through search and browse. Additionally, we needed to make the records not available through an OAI provider.

Our DSpace instance is 1.5.1, and we fortunately became aware of a patch for this version of DSpace, created by DSpace developers at Johns Hopkins University. Although our needs were different from theirs, and the patch required a good deal of customization to work with our DSpace instance, the patch provided a great starting point from which to begin. We had to add some additional code to ensure that the metadata for the records were completely hidden, and to ensure that browsing and searching would not be affected by the embargoed records. We also added a feature that allowed a user to change an embargo period for an existing record, rather than simply adding or removing it. We decided that we would use a preset number of embargo periods, including an option for an embargo period of 200 years for items for which the embargo period was indefinite. This 200 year option will become the default option for dissertations which are newly ingested from the Drop Box Processor. Once the appropriate permissions are granted for that dissertation, we can change the embargo period to make the record public, or to set the embargo to a period agreed upon by the submitter.

We have discovered that the embargo feature may have other uses beyond ETD's. Several campus groups that are interested in depositing content into IUScholarWorks have expressed interest in the embargo feature. Indeed, the embargo feature has been a much-requested feature by the DSpace community in general, as evidenced by its inclusion (in a somewhat basic form) in DSpace 1.6.

The Third Challenge: Library Catalog Integration

We have already implemented the Drop Box Service, and work on the embargo feature is nearing completion. Once it is complete, we can begin to ingest dissertations into our production DSpace instance. We expect that this work will definitely be completed by the time of the conference in July. There is one additional challenge that we will need to overcome, although it is not necessary to begin ingesting content. Our online library catalog, IUCAT, has a record for each IU dissertation. The library's

Technical Services department receives information from ProQuest about each dissertation, and creates a record based on this information. This workflow will happen in parallel to our own ingestion of the dissertation into the DSpace repository; the two workflows are not necessarily performed in a certain order.

Once the record for a dissertation has both been ingested into IUScholarWorks and is made available via IUCAT, a couple of things need to happen. First, an additional URL link to the DSpace handle needs to be made available in IUCAT. Secondly, the DSpace record metadata needs to be overlaid with the richer metadata available in IUCAT.

A number of solutions have been suggested for this problem, and we are currently considering them. There are a number of ways, both manual and automatic, that this data can be exchanged, and we hope to have a solution to this challenge soon. We are very interested in recent efforts to provide a mechanism for batch metadata updates, since this was another requested feature that was added in DSpace 1.6. This sort of batch update feature may prove very useful for our efforts.

Conclusion

The addition of ETD's into our DSpace repository presented a number of obstacles, and finding solutions to these issues proved to be a challenge. However, the experience has been rewarding. Not only have we been able to find solutions that work for us, but these solutions have been able to help us expand IUScholarWorks and DSpace in more general, reusable ways. We believe that these challenges may be common to other universities as well, and we hope that this presentation will provide insight to others who may be experiencing the same challenges.