Jul 28th, 1:55 PM

Storage Made Simple: Preserving Digital Objects with bepress Archive and Amazon S3

Lisa A. Palmer
University of Massachusetts Medical School

Storage Made Simple:

Preserving Digital Objects with bepress Archive and Amazon S3

University of Massachusetts Medical School
UMASS

Lisa A. Palmer
Lamar Soutter Library
University of Massachusetts Medical School

Hi, I'm Lisa Palmer from the Lamar Soutter Library here at UMass Medical School.

About a year ago we implemented the new bepress Archive service, which works with Amazon's Simple Storage Service (usually referred to as S3) to provide a real-time archive of repository content and metadata. My presentation will briefly describe this project.

Image credit: http://www.publicdomainpictures.net/pictures/110000/velka/rows-of-galvanized-buckets.jpg, License: CC0 Public Domain

In 2015-2016 my library went through a strategic planning process and developed a "purposeful pathway" to guide us for the next few years. One of the objectives that was identified was this one on the slide: to "responsibly preserve…"

Why did this become a focus? We know that bepress has a secure infrastructure with multiple backups. For years we received a quarterly archive from bepress that I dutifully downloaded to a network drive, just in case we needed local access to our files. But as our publishing activities and the amount of unique and born-digital content in the repository increased, we wanted to have more access and control over this investment: multiple journals and conference proceedings that we publish, as well as datasets, the version of record of dissertations and other student materials, etc.

We also wanted to follow best practices. The ability to preserve our content with a trusted and reliable external party is an important factor for journals that apply to be accepted into the Directory of Open Access Journals, and to be indexed in Scopus, MEDLINE, Web of Science and other content indices.

Even as a smaller institution, we knew we could and should be doing more - that we needed a long-term strategy.

Considerations

- Repository size (34 GB in June 2016)
- One solution for all content: journals, ETDs, e-books, data, proceedings, posters,...
- Restricted resources: staff, expertise, infrastructure, budget
- Ease of use

See: POWRR (Preserving Digital Objects with Restricted Resources) Project, http://digitalpowrr.niu.edu/; COPTR (Community Owned Digital Preservation Tool Registry), http://coptr.digipres.org/Main_Page; From Theory to Action: Good Enough Digital Preservation for Under-Resourced Cultural Heritage Institutions

3

We embarked on a project to compare digital preservation tools and services. The links at the bottom of the slide are useful resources if you're embarking on a project like this.

We wanted one solution for all our content in the repository, if at all possible. We don't have a large library staff, or library programmers, so we needed a service that was hosted and did not require substantial staff resources for programming, system administration, metadata provision, or ingest. Cost was another important factor, given the tight library budget.

We couldn't find a solution that was a good fit for us - services were too costly, or required expertise or infrastructure that we didn't have, or the business model didn't include preservation of non-journal content.

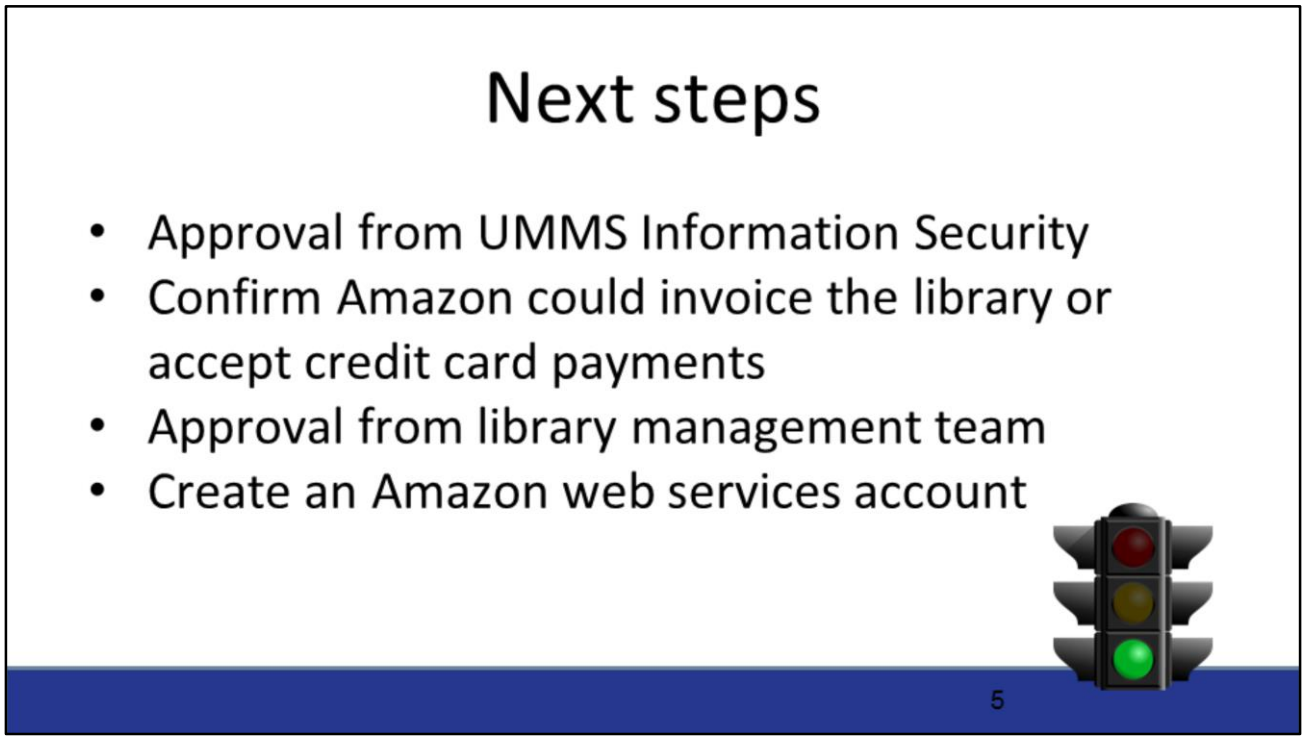
bepress Archive with Amazon S3

bepress, "Getting Started with bepress Archive,"
https://www.bepress.com/reference_guide_dc/getting-started-bepress-archive/

4

bepress released their Archive Service in May 2016 and I was immediately interested. This sounded like the solution we were looking for:

- real-time archive
- content and metadata are immediately PUSHED to Amazon S3, one of Amazon's cloud storage services, by bepress - I wouldn't have to upload content
- all objects in our repository would be preserved, and our SelectedWorks files and metadata as well
- industry-standard platform that is secure and reliable [Amazon provides redundant storage (data is replicated across multiple servers in geographically dispersed data centers within a region), performs data integrity checks and checksums to repair any corruption]
- web-based - no server requirements
- cost
    - on the bepress side, no additional fee or license for bepress subscribers
    - For Amazon web services, pricing is based on usage - it is not a subscription
    - Storage and transaction costs are inexpensive: https://aws.amazon.com/s3/pricing/. We estimated that S3 storage and transactions would cost only $36 per year. (AWS Simple Monthly Calculator.)

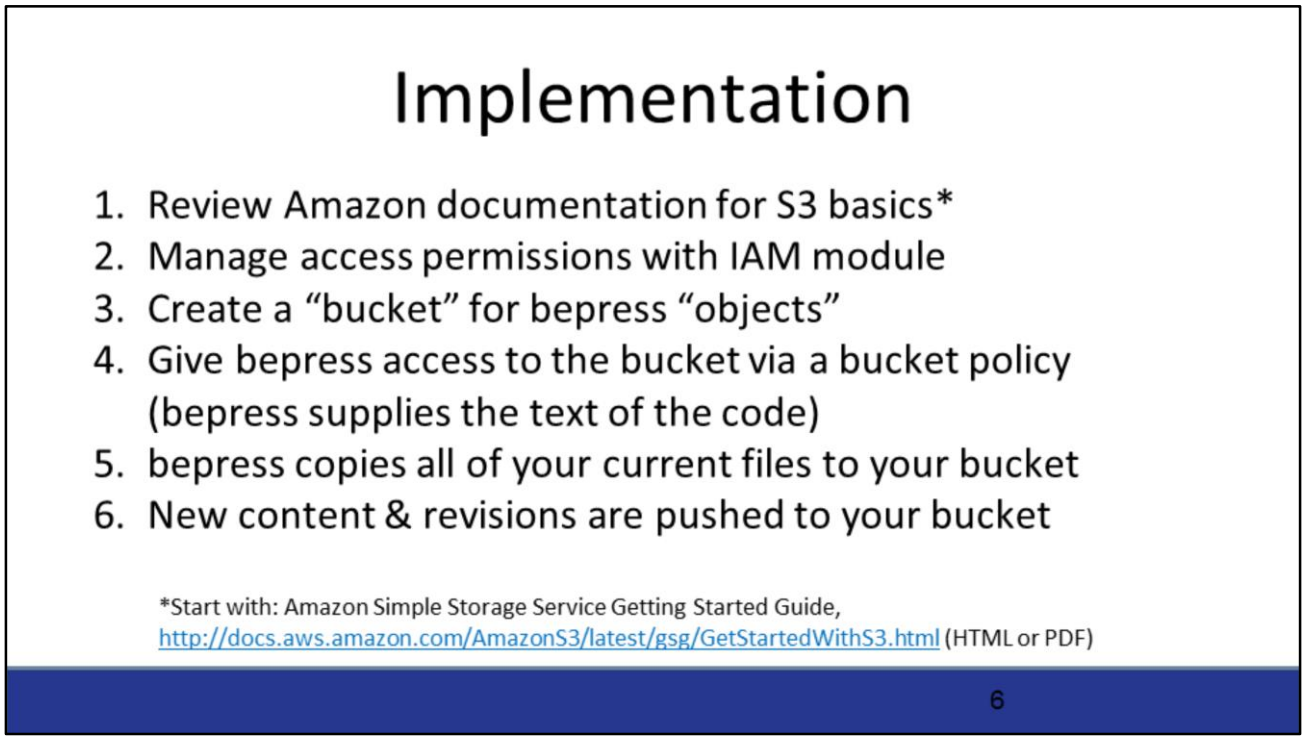


My department head Rebecca and I were convinced that this solution was the way to go. But we had a few next steps before getting approval from the library management team.

We first got approval from our campus' Information Security Officer to confirm that we weren't violating any institutional policies regarding the use of cloud services. We were fine because the data in our repository is all public information, classified as low sensitivity (not confidential).

Once we had IS security approval, and confirmed that we would be able to pay Amazon by credit card or by invoice, the library management team was on board and we were given the green light for me to get an Amazon web services account (easy online process)

Image: https://pixabay.com/en/green-light-traffic-light-signals-24178/, CC0



Then I had to learn how to use Amazon Web Services, which was easier than I anticipated. They have a large user base and their documentation is excellent, with plenty of examples. We've had good experience with their customer service. I set up 2 groups of users -- Administrators and Finance -- to give staff members access to the account information they needed. This way I could grant other people access to the account without having to share my root administrator password.

The basics are pretty straightforward:

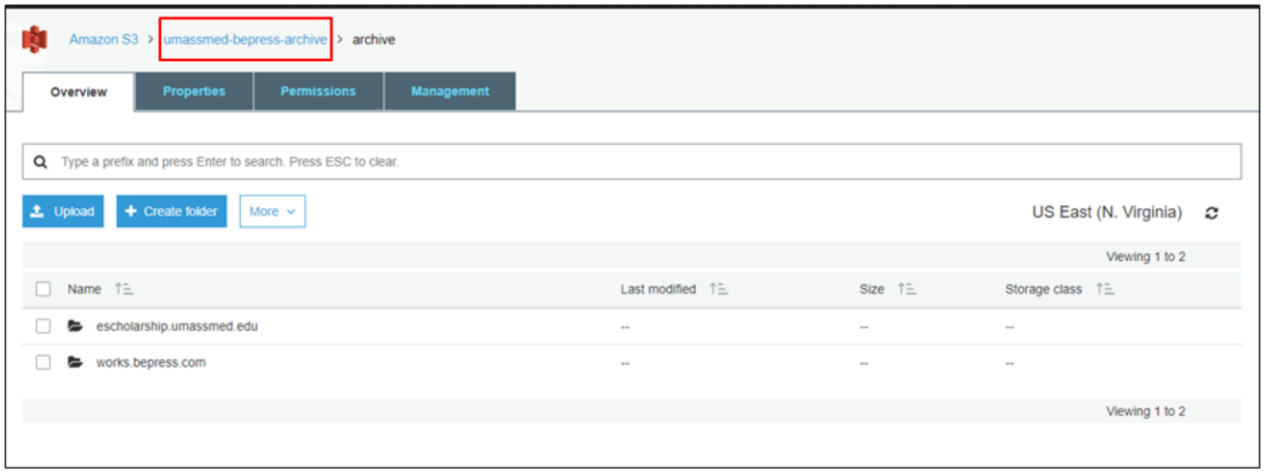
- S3 stores data as "objects" within "buckets"
- An object consists of a file and any metadata that describes that file
- Your repository and SelectedWorks files are all in 1 bucket

I created a bucket for our bepress content and then gave bepress access to the bucket using a snippet of code they gave me.

- Bepress copied all of our current files to our S3 bucket.
- Subsequently, new posts or revisions to the IR trigger an action to push this content to our S3 bucket within minutes. (I also set up a bucket to log all these transaction requests.)
- The file structure / hierarchy on bepress' server is replicated on Amazon S3 so that I can easily navigate to a specific file.

My understanding is that bepress does not delete content that has been removed from the repository, which means I would have to delete it manually from S3. I have not had occasion to need to do this yet.

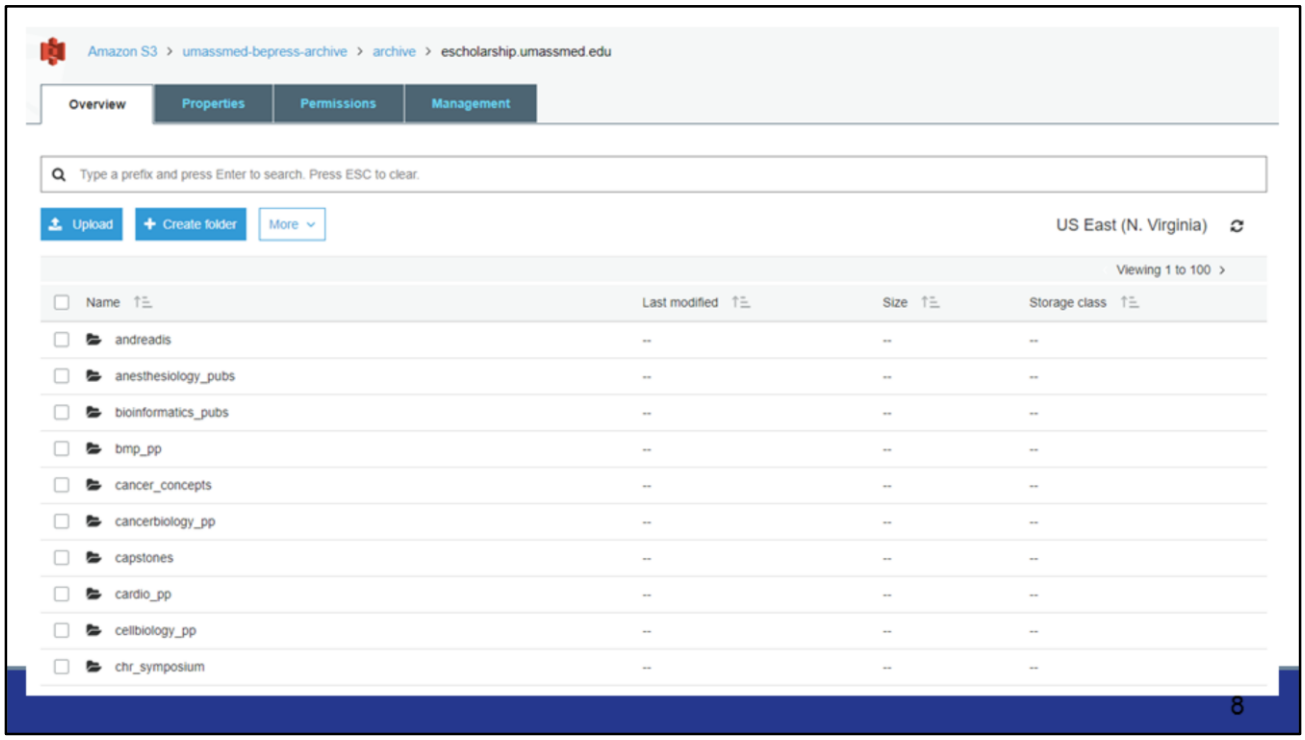
Our bepress bucket



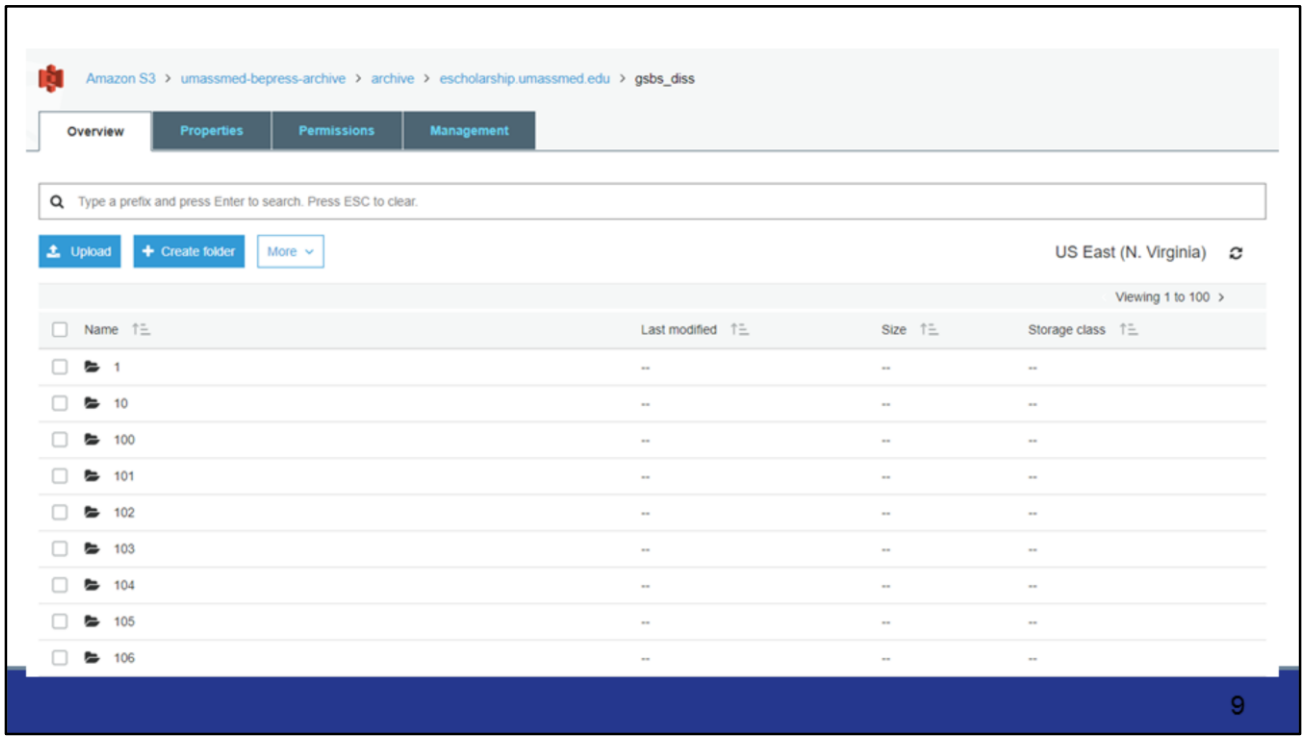
7

So what does the archive look like? Here are some screenshots from our account.

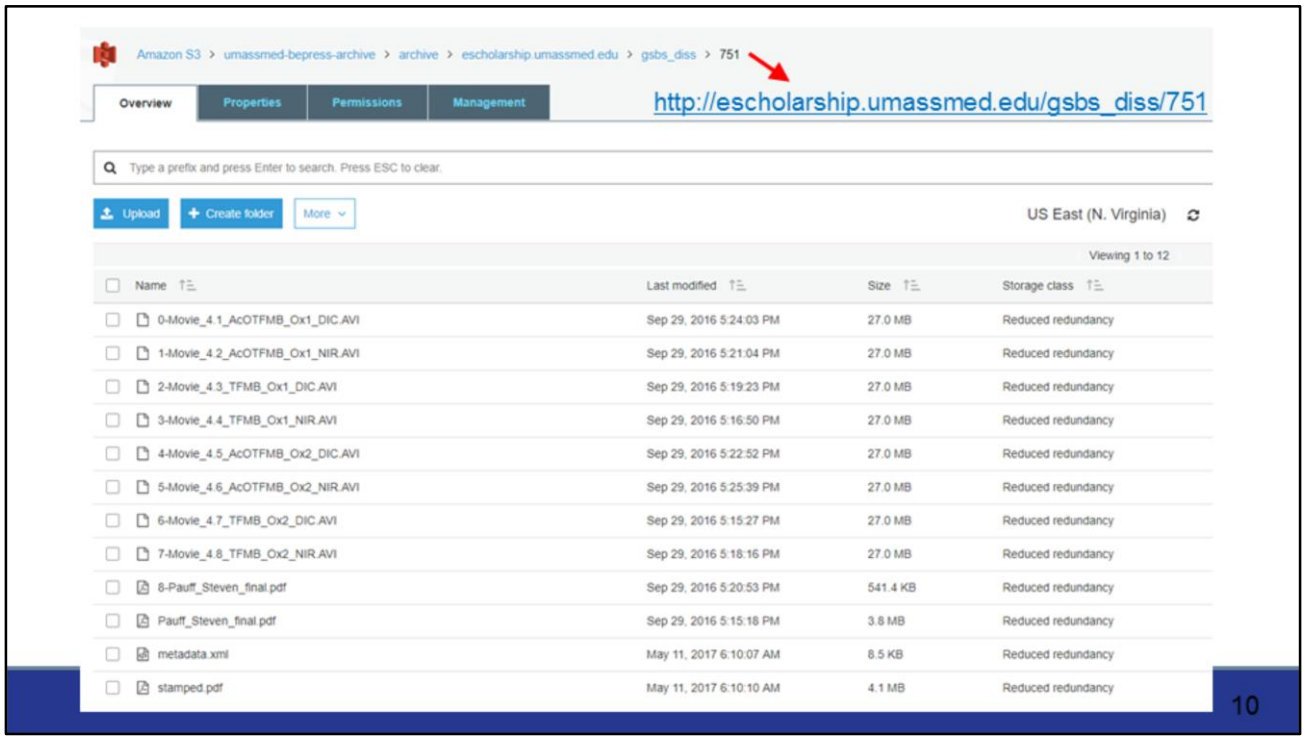
Here is our umassmed-bepress-archive bucket with 2 folders: 1 for our repository and 1 for content on our SelectedWorks profiles



If I click into the repository, you can see the folder structure with our series listed, sorted alphabetically



Here is one of our dissertation collections - the numbers correspond to the number in the URL, i.e. folder 1 is for http://escholarship.umassmed.edu/gsbs_diss/1



Here's dissertation 751:

- latest native file (there isn't one in this example since it was uploaded as a pdf and not a MS Word file)
- latest unstamped pdf (without download cover page)
- the stamped pdf (with download cover page)
- the metadata in xml format
- supplemental files (8 movie files that display publicly & 1 scanned permission form which is hidden)

metadata.xml


```
<?xml version='1.0' encoding='iso-8859-1' ?>
<documents><document>
<title>Advancements in the Synthesis and Application of Near-Infrared Imaging Reagents: A Dissertation</title>
<publication-date>2015-01-23T00:00:00-08:00</publication-date>
<authors>
<author>
<email>steven.pauff@umassmed.edu</email>
<institution>University of Massachusetts Medical School</institution>
<lname>Pauff</lname>
<fname>Steven</fname>
<mname>M.</mname>
</author>
</authors>
<subject-areas>
<subject-area>Dissertations, UMMS; Spectroscopy, Near-Infrared; Coloring Agents; Diagnostic Imaging; Esterases; Fluorescein;
Fluorescence; Fluorescent Dyes; Indicators and Reagents; Light; Optical Imaging; Oxazines; Rhodamines</subject-area>
</subject-areas>
<keywords>
<keyword>Near-Infrared Spectroscopy</keyword>
<keyword>Coloring Agents</keyword>
<keyword>Diagnostic Imaging</keyword>
<keyword>Esterases</keyword>
<keyword>Fluorescein</keyword>
<keyword>Fluorescence</keyword>
<keyword>Fluorescent Dyes</keyword>
<keyword>Indicators and Reagents</keyword>
<keyword>Light</keyword>
<keyword>Optical Imaging</keyword>
<keyword>Oxazines</keyword>
<keyword>Rhodamines</keyword>
</keywords>
<disciplines><discipline>Biochemistry</discipline>
<discipline>Chemistry</discipline>
<discipline>Molecular Biology</discipline>
</disciplines><abstract>&lt;p&gt;Fluorescence-based imaging techniques provide a simple, highly sensitive method of studying live
cells and whole organisms in real time. Without question, fluorophores such as GFP, fluorescein, and rhodamines have contributed
vastly to our understanding of both cell biology and biochemistry. However, most of the fluorescent molecules currently utilized
suffer from one major drawback, the use of visible light. Due to cellular autofluorescence and the absorbance of incident light by
cellular components, fluorescence imaging with visible wavelength fluorophores often results in high background noise and thus a low
signal-to-noise ratio. Fortunately, this situation can be ameliorated by altering the wavelength of light used during imaging.
Near-infrared (NIR) light (650-900 nm) is poorly absorbed by cells; therefore, fluorophores excited by this light provide a high
signal-to-noise ratio and low background in cellular systems. While these properties make NIR fluorophores ideal for cellular
imaging, most currently available NIR molecules cannot be used in live cells. The first half of this thesis addresses the synthetic
difficulties associated with preparing NIR fluorophores that can be used within living systems. Small molecule NIR fluorophores are
inherently hydrophobic which makes them unsuitable for use in the aqueous environment of the cell. Water-solubility is imparted to
```


Here's the metadata for that dissertation. It's not in OAI-PMH or standard format. The elements correspond to field names in Digital Commons. This is just the beginning of the file - the entire file is quite long and includes the URL, description, and mime type for supplemental files. NOTE: Hidden supplemental files are not currently included in the metadata.

You can drill down to SelectedWorks content in the same way. In our case most of those folders contain only an xml metadata file because the objects themselves are stored in the repository.

Total $7.29 for 10 months (for the initial ingest + 2400 new works posted + revisions to objects and metadata)

We feel that bepress Archive with Amazon S3 is an excellent, easy to use digital preservation and electronic archiving option. It's a good solution for us at this time and we're pleased to have this increased preservation. Looking ahead, we might be able to leverage our S3 account to preserve additional content, perhaps from our archives.

Thank you!

Lisa Palmer, Lisa.Palmer@umassmed.edu

@lapalmer14

https://works.bepress.com/lisa_palmer/