Biz of Digital — Developing and Growing a New Repository Service: Part 2 Procedures for Library Submissions

Michelle Flinchbaugh

Biz of Digital — Developing and Growing a New Repository Service: Part 2 Procedures for Library Submissions

by Column Editor: **Michelle Flinchbaugh** (Acquisitions and Digital Scholarship Services Librarian, Albin O. Kuhn Library & Gallery, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; Phone: 410-455-6754; Fax: 410-455-1598) <flinchba@umbc.edu>

Column Editor's Note: *This is Part 2 of a 3 part series on Creating a New Repository Service. Part 1: Getting Started appeared in the June issue (v.31#3). Part 3: Expansion will appear in the November 2019 issue.* — *MF*

Introduction

The **University of Maryland, Baltimore County (UMBC)**, a research-intensive institution with 546 Full-time and 292 Part-time faculty, participates in the Maryland Shared Open Access Repository (MD-SOAR) and has the MD-SOAR DSpace platform available to it. The Digital Scholarship Services Librarian (DSS Librarian) at **UMBC's Albin O. Kuhn Library** shifted duties from Acquisitions to work full time on the repository. Four months into a soft roll-out with minimal outreach only to individual faculty members, she had attempted to get faculty to submit items themselves with little success. She also began identifying new faculty publications via Google Scholar Alerts and adding these to the repository when appropriate with this method of populating the repository proving to be far more successful.

Processing Google Scholar Alerts requires determining if items were appropriate for the repository, checking rights, asking faculty for the item or a particular version when needed, then adding the work. Sometimes when the librarian corresponds with faculty about their works, they also request that she load other materials as well. After four months of this approach, the librarian was inundated by requests to add items to the repository for faculty, so she further developed processes and procedures to handle those requests as well.

With new procedures being developed primarily by one librarian, it became critically important to document in detail both so that items would be entered consistently and so that another person could find and follow the procedures for processing and submitting items to the repository for faculty. In approaching the development and documentation of procedures, there were a number of questions that arose:

- Given the MD-SOAR scope that requires items be available for free, either via a link to a free version online or via an attached pdf file, and **UMBC's** policy decision to add items to all relevant collections, including Student, Faculty and Staff Collections, what are the steps for processing an item on a Google Scholar Alert?
- How does processing vary when the works don't come from a Google Scholar Alert, but another repository, from a publications website, a Google Scholar Profile, a CV or from a list?

Processing Google Scholar Alerts

Creating and De-Duping Google Alerts

Results for any search performed in Google Scholar includes the option to "Create alert." When the searcher chooses "Create alert," the terms of the search fill into an "Alert query" box, and the searcher's email auto-fills into another box. Once the searcher clicks "Create alert," she receives an email of new items whenever new items with that search term are added. The DSS Librarian chose to monitor the search terms "The University of Maryland Baltimore County" and "UMBC." The two Alert emails come with many duplicates, and the first step is to print both the **UMBC** and **University of Maryland, Baltimore County** Alerts that came on a given date, and remove duplicates by crossing them out on one of the Alert printouts. At first it was unclear if both printouts are necessary. Indeed, most **UMBC** publications come on the Alert with the full name, **University of Maryland, Baltimore County**. Yet there are still consistently unique **UMBC** publications on the Alert for the abbreviated form of the university name, especially for preprints and other informally published items where the full name of the university wasn't included on the work.

Figure 1 Google Scholar Alert Example

Determining if the Works on Alerts are Appropriate

About half the items the DSS Librarian receives on Google Alerts are inappropriate to the repository, already loaded into the repository, or not **UMBC** publications at all. Some items are only abstracts without the full work, CVs, obituaries, patent applications, or a description of a grant funded project. These aren't added to the repository, so they're crossed out on the printout. Google Alerts include theses and dissertations, but they have their own separate workflow and are periodically loaded, so not processed when received on a Google Scholar Alert.

Some items have no **UMBC** author, but include **UMBC** in a citation or credit, or **UMBC** stands for another organization. Therefore, the first step in processing Google alerts is to determine if the format of the item is appropriate, if it's appropriate to add it via this workflow (it's not a thesis or dissertation), and that at least one author is affiliated with **UMBC**. Since these are new publications, they aren't generally duplicates, but in instances where the title sounds familiar, the DSS Librarian also searches the repository to see if the item has already been added. Otherwise duplicate searching is done right before items are entered.

Paywalls

Once the DSS Librarian determines that an item is appropriate for the repository, she notes on a printout of the email alert if the item is paywall protected. When she's asked faculty for permission to load a work in *ArivX.org*, or a pre- or post-print pdf to add, a month or two later she follows up on works that she's requested but not received a response. Works not paywall protected are free and can be added with just a link if permission wasn't granted for the file or a pdf wasn't provided, so she adds these without a file attached once it's clear that no response is ever coming. Paywall written next to an item indicates that the item can't be added unless the faculty member granted permission or provided an appropriate version because it's not available for free, so no follow-up is necessary.

Determining Collections

Collections in *ScholarWorks@UMBC* include both departmental collections, e.g., **UMBC** History Collection, **UMBC** Physics Collections, and author status collections, e.g., **UMBC** Student Colletion. Determining collections may be done early in the process, or at the end. If the DSS Librarian is searching the directory to determine if an author(s) is affiliated with **UMBC**, or searching for author(s) email address, she'll do it while already in the directory. Often the department(s) of the **UMBC** author(s) is given on the work, and the item will go in each of the collections for all departments listed as affiliations for that author. Then she searches the **UMBC** Directory to determine if the author(s) are faculty, staff, or a student, and indicates whichever one(s) are appropriate. She also uses the campus directory to find the departmental affiliation of authors when it's not given on the work.

Making determinations of what collections an item belongs in is sometimes hampered by the limitations of the **UMBC** Directory. Those who have graduated or otherwise left the university are no longer listed. Generally, the directory explicitly states that someone is faculty and their department. Sometimes it also explicitly states that a person is a graduate assistant and their department. For graduate students who aren't graduate assistants, and for undergraduates, there is no information on a person's status or affiliation, but status can sometimes be determined from the department's website. Occasionally there is no information to determine a person's status or affiliation. If the DSS Librarian has an email address for the author (either given on the work or via the e-mail system), she'll ask the author. Items can only be mapped to collections if a determination on status and affiliation is possible based on the available information. If both the status and affiliation cannot be determined, the item cannot be added to *ScholarWorks@UMBC* because it must go into at least one collection.

If making the collection determination early on in the process becomes problematic or time consuming, the DSS Librarian generally puts it off until the end so as not to be overly distracted from the steps she's currently working on, and because she might not be able to add the item, and won't actually need to make collection determinations.

ArXiv

Many **UMBC** faculty consistently post pre-prints on ArXiv, and the DSS Librarian receives notices of all of **UMBC** posts to ArXiv through Google Scholar. ArXiv allows reposting to institutional repositories with the author's permission, so the DSS Librarian emails the author, with the title of their work in the subject line, to seek permission. She uses a canned request in a Google template for this.

> Security/Privacy Attitudes and Unlock Authentication, in ScholarWorks@UMBC?|
>
> Hi Sebastian,
>
> UMBC recently implemented a digital repository, ScholarWorks@UMBC, here: https://mdsoar.org/handle/11603/14. We're trying to add more faculty publications so that ScholarWorks better showcases what we do at UMBC.
>
> I'm wondering if you'd give me permission to put your pre-print paper recently deposited to ArXiv, Toward Understanding Connections between Security/Privacy Attitudes and Unlock Authentication, in ScholarWorks@UMBC?
>
> If your ok with it also appearing in ScholarWorks@UMBC, simply reply yes. Thanks.

Figure2 ArXiv email example

Response to these emails has been very high. Some faculty have become "regulars" always saying yes. Once they become a "regular" she omits the explanation and niceties and just sends a single sentence question asking if it's ok. With a positive response, she proceeds with adding the pdf of the item. With a negative response, she adds the item, linking to it on ArchivX without providing the pdf. Finally, with no response, after a period of time has elapsed she also adds the item to the repository, linking to it on ArchivX without providing the pdf.

Rights — Creative Commons Licenses

If the item is not on ArXiv, the next step is to investigate rights for the item. First she looks for a **Creative Commons** license on the work, and if there is one, she adds the item to *ScholarWorks@UMBC* on the same **Creative Commons** license. If the item says open access on it, this requires some investigation, as sometimes it means everything in a particular journal is on the same **Creative Commons** license, and other times they've defined it in a particular way. Oftentimes she can load these into the repository, but sometimes the publisher means only open on their website and doesn't allow distribution via a repository, in which case she simply adds the item with a link to the publisher's version without providing a file.

Discovering that an item is on a **Creative Commons** license sometimes doesn't happen until later in the process. If a publisher is completely open access, they may provide information on their website about what **Creative Commons** license all of their journals are on, but not provide that information on the journal page or on individual articles. When this is the case, we may not realize an item is on a **Creative Commons** license until we see that information in a reference such as our Policy of File or in the Sherpa-Romeo database.

Rights — Federal Government Publications and Federal Government Employee Authors

Part of investigating rights is determining if the item is a U.S. government publication, or a work authored by a U.S. government employee as a part of their job. If so, these items are in the public domain, so she adds them to the repository, putting them on a **Creative Commons** Public Domain license, adding a note that states, "This is a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law" or "This work was written as part of one of the author's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law."

Rights — Our Policies on File and the Scherpa-Romeo Database

If a work isn't on a **Creative Commons** license, or in the public domain, and it's a journal article, the DSS Librarian finds the journal's

self-archiving policy using a "Policies on File" document (*https://wiki. umbc.edu/display/library/Policies+on+File*) or the Scherpa-Romeo database of publisher copyright policies and self-archiving (*http://www. sherpa.ac.uk/romeo/index.php*). Initially she used only the Romeo-Scherpa database, but conference proceedings publishers aren't included, and as these became quite numerous she created a "Policies on File" document for them. In this document, she keeps a summary of each publisher's policy that includes all terms to address, and a link to the full policy.

Eventually she added all publishers' policies she frequently uses to the "Policies on File" document, and additionally, policies that were emailed to her by the publisher. By including frequently used publishers in this document, she saves time in that she doesn't have to search for policies for those publishers, which sometimes takes a great deal of poking around on their website. Additionally, it saves time in that she doesn't have to read complex, confusing, or lengthy legalese in policies and author agreements to find the information she needs each time she has another work from that publisher. Eventually, it also provided a way to enable students and staff to make decisions without them also having to review such complex and confusing documents and agreements. By including policies emailed to her by the publisher, she retains a record of what she was told so that she doesn't have to ask each and every time she receives a work from that publisher. The Policies on File document is available on the **UMBC Library's** intranet with other documentation, here: *https://wiki.umbc. edu/display/library/Policies+on+File*. It will be important to periodically check this document against the publisher's policies to note any changes that have been made and to ensure that links are still working.

The scope of Google Alerts is journal articles and case law, so the initial procedure for finding rights information assumed that all works retrieved via Google Scholar Alerts were journal articles, but they also include conference papers, presentation slides, book chapters, reports, unpublished items etc. To address this, she systematically focused on only journal articles and conference papers, both of which were available in substantial quantities. The first step for these was to check them against the "Policies on File" document, frequently adding publishers' policies as she receives quantities of their works. If the DSS librarian doesn't get the self-achiving information there, what she does varies based on whether the item is a conference proceeding or journal article. For conference proceedings, she searches the conference, conference website, and conference proceedings to look for a posted policy. If she can't locate a posted policy for a conference, she'll link to them without providing a pdf if the content is available freely on the web; if it's not freely available, that one will be skipped. If there is a large quanitity of works from the same conference, she'll look for a contact and ask about their policy. For journal articles, she searches the Scherpa-Romeo database, and if there's no information there, she'll search for the journal or journal publisher and try to locate a self-archiving policy on their site. When trying to find self-archiving policies for either conference papers or journal articles, the publishers don't usually call them that, so it often takes some poking around on their website and perusing a few different pages before finding the one(s) with the information needed.

When the DSS Librarian locates rights information, she sometimes discovers that the entire conference proceedings or journals are open-access, or everything that publisher publishes is open access, even though this isn't indicated on the work itself. She has to determine if by open access they mean that it's on a **Creative Commons** license, free on their website, or are using some other definition of open access. It can also take some searching on the publisher's website to find the **Creative Commons** license the proceedings or journal are on. If she has confirmed it is on a **Creative Commons** license, then she handles as described above under "Rights-Creative Commons Licenses." If

open access is defined some other way, she stays within the publisher's definition of it.

Sometimes the DSS Librarian learns from Scherpa-Romeo that she can post the published version, and if that's accessible to her, she goes ahead and adds it. If it's not accessible, she asks the author for it. Most frequently she finds in Scherpa-Romeo that only the pre-print or post-print of an article can be posted. These are also known as the submitted and accepted version of the article. The pre-print or submitted version is the manuscript before peer-review took place. The post-print or accepted version is the manuscript after the author has made peer-review edits, but before the publisher has done any work on it. To get these versions of an article, she generally has to email the author to request them. Her first email about this simply stated she was trying to add items to a new repository. This resulted in a decent number of responses, but a second email that focused on how much open access can increase the citations to works only available for a fee improved the responses significantly.

> Providing open access (free access) to articles published in subscriptions journals increases citations by 25%-250% (mean 96%, media 73%). UMBC recently implemented a digital repository, ScholarWorks@UMBC, here: https://mdsoar.org/handle/11603/14 for providing open access to UMBC publications.
>
> If you'd like to extend access and the impact of your recently published work, "Igloo Historiography parts I, II, and III," I'll put it in ScholarWorks@UMBC for you. Your publisher, the History Compass, allows the submitted (before peer review) or accepted version (after peer review but before the journal has done any work on it) to be placed in institutional repositories, so if you're wiling, please send me one of those version. Please let me know which version you send.

Figure 3 Asking for a pre-print or post-print with explanation of how open access increases citations

If an author sends the pre-print or post-print, she adds it to the repository. In the event the author provides the published version rather than the pre-print and post-print, the DSS Librarian explains the risk of copyright infringement and provides more detail on the version she actually needs; sometimes they respond with an appropriate version and sometimes they don't. If they don't respond, or don't provide an appropriate version, she addresses the work during the follow-up process described above under "Paywall."

Google Scholar Alerts procedure documentation is available here: *https://wiki.umbc.edu/display/library/Google+Alerts*. The DSS Librarian updates it as time permits.

Rights — Terms

Unless a submission to the repository is only a link or unpublished, specific terms must be adhered to. The terms of all **Creative Commons** licensed materials require that a citation be included, and the inclusion of a citation is so ubiquitous that it can be assumed to be required on all published works. Most publishers also require a link to the final published version of the work, and/or a DOI linking to the final published version of the work. Some require copyright statements, and some require specific statements with information on the work, sometimes the full citation, plugged in. Finally, some require embargo period be adhered to, a period of time after official publication and before which the work can be made available via repositories. Care must be taken to note and adhere to the specific terms each publisher requires for inclusion in the repository.

Processing Requests to Load Materials

When the DSS Librarian contacts faculty with a request related to content discovered via Google Scholar Alerts, they occasionally respond with their own requests for additional materials for the repository. The first of these types of requests came from faculty that had come from another university that had a repository, and they wanted their materials from that repository added to ours. Sometimes when faculty respond to our requests for a work, they attach the pre-print or post-print of other papers that they've written. Other times, they've directed us to their public Google Scholar Page, their Lab publication page, or a facility publication page. Sometimes they send a list of everything that they've published or their CV. Later in outreach, the

Digital Scholarship Services Librarian would tell them that they can send lists, CVs, or a link to their Google Scholar profile and the library would process and add these items.

The additional publications that the DSS Librarian receives are handled exactly in the same manner as she handles the Google Alerts, with some key differences:

Variations in the Materials and Information Provided

PDFs: When working from Google, there is usually a link to the full-text version of a work on the publisher's website, or to the publisher's record for the item with the link to the work. Most often this isn't a version the DSS Librarian can load and she has to ask one of the authors to provide the pre-print or post-print. The pdf on a local website (e.g., a lab or faculty webpage) may or may not be a version that can be loaded in the repository. Publisher's versions which usually can't be posted in repositories, are readily identified by the publisher's trademark, their copyright statement, and with pagination that doesn't begin with one. Pre- and post-prints can be identified by the lack of this information. When in doubt, the version posted can be compared to the published version. Depending on the age of the item in question, the DSS Librarian may or may not ask the author for a version to load. When working with some sources, she finds the versions to almost always be a version that can't be added, and with others, they've consistently posted a version that can be added.

Links: Google usually links to the record for a work, so metadata with a great deal of information and the published version of the work is instantly available. But sometimes Google links directly to the pdf, in which case the DSS Librarian searches to find a record with metadata because she wants to link to the published version, provide a DOI, and get metadata and a citation from the publisher's record. When a lab or faculty website has a link to a pdf of the full text, the lab or faculty website may or may not have a link to the published document — when it's omitted she searches for it because, again, she wants to link to the published version, provide a DOI, and get metadata and a citation from the publisher's record.

Metadata: Google usually provides accurate but limited information about a work, e.g., the work's title, the name of the journal it was published in, volume and number, pages, and publisher's record can be used to complete that information. On the other hand, when working with websites, lists, and CVs, titles don't always exactly match the title of the published version, or might include abbreviated or even erroneous journal information, making it difficult to locate the published version of the work. Generally, a title search on Google will yield the published version of the item. If it can't be located that way, the DSS Librarian searches the journal title or an abbreviated form of it to try to find the item on the journal's site. On the journal's site, she title searches, but sometimes when the items that aren't coming up by title, she'll also search by the author. Some journal websites don't have search capability and she has to navigate to the work by volume and issue.

Locating pre-prints and post-prints: The DSS Librarian seldom searches titles of works coming on Google alerts. First, she doesn't need to locate the publisher's record since there's usually a link to it, and also because publications are usually new, the work isn't usually posted on other sites yet. With websites, CVs, and lists of publications, she usually title searches items, both to locate the publisher's record if necessary, and also to look at the work on other sites, where she can sometimes find a free full text version of the article that she can either load or link to. For example, frequently she finds biology works available for free on PubMed and is able to load the version on PubMed or link to the version on PubMed.

Variations in the age of items: Google Alerts only provide notification for newly published items. The DSS Librarian also consults other lists of items published before the author was at **UMBC** which sometimes include items that were published more than 20 years ago. When a work is more than 20 years old, a publisher's current self-archiving policy certainly doesn't apply to it. Additionally, it is not always feasible to determine status and departmental affiliation of authors on older works because authors are more likely to have left **UMBC** and are no longer in

the directory. Therefore, unless it's freely available or on **Creative Commons** license, she does not do anything with items published that long ago.

For more recent content, if an item was written before the person came to **UMBC**, she adds the item on their department's page but doesn't add it to the Faculty Collection since the individual wasn't **UMBC** faculty when they wrote it.

Permissions: When the DSS Librarian discovers a **UMBC** publication via Google Alerts, she asks permission to load when the item isn't on a **Creative Commons** license and the author(s) own the copyright. However, when she receives a request to load the item, either individually or as part of list or Google Scholar page, permission to load is implied, so she doesn't ask.

Fewer out-of-scope works: When working from websites and CV's, out-of-scope works and works without an author affiliated with **UMBC** are extremely rare compared to Google Alerts and Google Profiles.

Information on status/department: When loading publications from a center or lab website, authors' profiles are sometimes available, precluding the need to search the directory for that information. When it's not on the work, and she sometimes get a lot more information from the website than she'd get from the directory if the department maintains historical records as opposed to deleting people when they leave **UMBC**.

Keywords for labs: When loading materials from a lab's publication page, she adds the name of the lab as a keyword. This allows for keyword searching that will generate a link to the lab's publications which can then be shared with the lab.

Determining How to Document Procedures for Different Types of Sources

While there is a lot of variation in the different sources of items to be processed, there are also enough key similarities and overlaps to make it preferable to manage a single long procedure, outlining when you do and don't perform certain steps. The DSS Librarian didn't begin this integrated document until a student began working on this. It's intended to eventually be a catch-all document that describes how to handle 90-95% of works received with any exceptions to be referred to the DSS Librarian for processing. This procedure will be discussed in more detail in the upcoming Part 3, Expansion, for this series.

Procedures for other formats weren't documented until a student was hired and began working on this, when the existing procedure and documentation proved inadequate in failing to provide information on anything but serials and conference proceedings. Those procedures will also be discussed in more detail in Part 3, Expansion.

Adding a Work to *ScholarWorks@UMBC*

Early on, when creating metadata records in *ScholarWorks@UMBC*, the DSS Librarian realized that she was handling some things inconsistently, so she documented what goes into each field (available here: *https://wiki.umbc.edu/pages/viewpage.action?pageId=73893118*). She would process an item, and then enter it. When doing this, she would have to note or remember information as she found it to fill in the metadata accurately. However, it was still easy to forget to add some bits of information, so she needed a consistent method of adding items that guided her through everything that she needed. Additionally, some of the information required judgement calls that a new student assistant or staff person wouldn't be able to make. Further, the text in the documentation was very dense, making it difficult for a student assistant or staff person to follow while actually entering new items in the repository. Most of these issues were resolved later in advance of hiring and are discussed in the next session.

Conclusions

Initial procedures and documentation for an operation relying on one librarian were an important stepping stone. While they weren't completely satisfactory, or always thoroughly documented, opportunities for changes were identified and made, in an iterative process. It allowed time to accumulate information, test, and revise. This made the procedures work better, and the documentation more complete and easier to follow. Having these interim procedures and documentation in place facilitated the expansion of the service described in Part 3 by allowing for intense focus on making the procedures and documentation complete, easily understood, and readily usable by student assistants and potentially eventually by staff. 🐾