Proceedings of the 8th Annual
Federal Depository Library Conference

April 12 - 15, 1999

PURLs: What Do I Need to Know? Working with PURLs in Your Local Catalog Background on PURLs and Link Maintenance in the Local Catalog

Arlene Weible, Willamette University
Salem, OR

Nan Myers and I are pleased to be with you today to talk about Persistent Uniform Resource Locators, also known as PURLs. As part of our work on the GODORT Cataloging Committee, Nan and I have been talking about PURLs for what seems like forever, but it's actually been just over a year. As a result of the work of the Cataloging Committee and the efforts of Tad Downing, Chief of GPO's Cataloging Branch, the depository community has been provided with quite a bit of information from GPO about their implementation of PURL technology.

What Nan and I would like to do today, however, is bring the discussion of PURLs into the library, and try to address some of the issues that need to be considered when working with PURLs in the local library catalog. We cannot promise to solve the particular problems that each library and library catalog system may face; however, we do hope to provide you with some information about the questions you need to ask, and resources that you can turn to when making decisions for your own library.

In order to get this discussion started, we do need to provide a little bit of background on PURLs and GPO's use of them. I'll begin with a definition, and then describe how GPO is using PURLs in their management of the FDLP Electronic Collection and cataloging activities. I will then discuss what PURLs will and will not do, which will lead me to some of the issues related to link maintenance in library catalogs. Nan will then take over and report some of the results of a survey she conducted on library procedures, and provide a checklist for local decision making.

Let's start with a definition. Described simply, a PURL is an actual URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. This service associates the PURL with the actual URL and returns that URL to the user. The information required to make this redirection possible is maintained in a record located in a resolution or PURL server. OCLC developed this technology in 1996, and was the first to implement its use with its Intercat project. I don’t have enough time to get into a more detailed explanation of the technology behind PURLs, but if you are interested, you can find more information about the development and technical details of the service on OCLC’s PURL Service Web page, available at <**http://purl.oclc.org**>.

While PURL technology had been around for awhile, it wasn’t until early 1998 that depository librarians found themselves face to face with the technology. This is when GPO first implemented the PURL Resolution Service to assist in the maintenance of URL information in the bibliographic records and Web pages they create. When resources are identified for inclusion on the Browse Electronic Titles Web page, or are identified for cataloging treatment, GPO staff create a PURL for the resource on the GPO PURL server. GPO utilizes an automated naming process, which assigns unique, consecutively assigned accession numbers to each PURL created.

Once the PURL record is created, the PURL is used as the link to the resource on the Browse Electronic Titles page, and is sent to the GPO Cataloging staff for inclusion in the resource’s bibliographic record. GPO catalogers add PURLs to the records currently created, and will add a PURL to existing records as they come up in the regular review process. Right now, GPO has no plans to systematically add PURLs to all records that contain URLs, although they have made an effort to add PURLs to all records for the resources listed on the Browse Electronic Titles page.

Another way to gain a better understanding of PURLs is to examine what they will and will not do. Let’s start with what they will do. PURLs may be used as a tool to maintain a constant link to an Internet resource, regardless of changes to the location of that resource. Once a PURL record is created, the PURL associated with the resource will remain constant. As the location of the resource changes, the PURL record may be updated to reflect the change, so the end user does not have to make note of the new location. Also, the PURL record retains the history of changes to the location, serving as a "travel diary" of sorts for a particular resource.

This is an example of a PURL record and the information it contains: <**www.willamette.edu/~aweible/dlc/purlrec.htm**>. You can see the PURL, the current URL, and the dates of maintenance. We can also see the different URLs this resource has had since it was first entered last July. As you can see, this resource has done a bit of traveling. One of GPO’s primary goals in implementing the PURL Resolution Service was to prevent the time consuming task of updating URL information in catalog records. The obvious advantage to using PURLs in catalog records is that records do not have to be edited each time a location changes, only the PURL record on the GPO PURL server needs to be changed.

PURLs clearly offer advantages, but there is an important issue that PURLs do not address. PURL records may only work to maintain a constant link to a resource if they remain up to date. Identifying Internet resources that have changed location remains a challenge. GPO has made a commitment to maintain the PURLs they create with current URL information. This is accomplished by the use of OCLC software that provides the ability to check for valid links. GPO runs this software on a weekly basis to identify broken links. Because this is an automated process, however, it has its limitations. For example, while software may be able to report that a URL exists, it cannot determine if the content of the resource remains the same.

Maintenance of PURLs, while automated to some degree, still requires human intervention. GPO has reported that they currently use at least 2 FTE to maintain accurate records in the PURL server. They are assisted by reports of incorrect URLs/PURLs via the askLPS service. In fact, it is vitally important that the automated process of link validation be supplemented with human oversight to assure that the PURL Resolution Service remains effective.

At this point, I should also mention the role of catalog record vendors in the process of keeping links up to date. While it is true that vendors are making efforts to keep the links valid in the records they distribute to libraries, it is my understanding that they are not using a systematic process, such as automated link validation software, to accomplish this task. While they accept reports from users who identify broken links, they primarily rely on the maintenance activities of GPO to keep the records up to date. I think it is important that libraries using vendors for the delivery of GPO catalog records verify exactly what the vendor is doing to keep links valid, and determine if their efforts are sufficient for the needs of your local library.

Link maintenance is an important topic I'd like to take some time to discuss. While it is clear that GPO has taken seriously its commitment to keep PURLs accurate, I feel very strongly that libraries must also undertake link validation activities. Since GPO is not systematically going back to every catalog record it created to add a PURL, it is going to be awhile before all GPO records have PURLs. And, because it is such a large task, GPO also needs help with maintaining accuracy in PURL records. In my opinion, libraries must also do what they can to check not only the existing URLs, but also PURLs. This will help to ensure that the links provided in library catalog records remain viable access points to Internet-based information resources.

If the links aren't valid, users won't consider the catalog to be a good access tool, and all this effort to provide links from the catalog will be wasted. So, to help encourage you, I'd like to show you what my library does to maintain accurate links in our catalog. I will also discuss some of the issues that must be considered when working with PURLs in this process.

Let me start with just a few words about our library. We have a depository item selection rate of about 25%. The process I am about to describe is done on a monthly basis, with the help of the library's systems assistant.

The first step is to extract the records that contain URL/PURL information in the 856 field. We have to do this because our system, Innovative Interfaces, does not currently have an automated link checking component. It is my understanding that they currently have a program in beta test, but until this is available, we have developed an interim procedure that works for us.

Using our system's list making features, we are able to create a file of records that can be exported from the catalog. The next step is to convert this file to HTML format. We are indebted to Tom Tyler from the University of Denver, who has created what he calls MARC-X-GEN software to help convert this file of MARC records to HTML. This has saved a tremendous amount of time, since the previous method we used involved a lot of manual editing. This software was designed to work with files generated from Innovative systems, but Tom has indicated that the software should work with files from any system, as long as the records are in MARC format.

You can find more information about the MARC-X-GEN software in Tom's paper on maintenance issues in the Web-accessible OPACs <**www.du.edu/~ttyler/cil99/proceedings.htm**>. The file that is generated looks something like this <**www.willamette.edu/~aweible/dlc/purl.htm**>. The software converts the URL or PURL into an active HTML link, making the title of the resource the text. It also records the OCLC number, and additional note information contained in the 856 field.

Once this file is created, we use a software program called LinkBot to check the file to verify the validity of the URLs and PURLs we've exported from the catalog. As far as choosing this software over the other link validation programs available, I have to say that I relied on the expertise of my library's systems staff. They were already using this software to maintain the links in the library's Web pages, and it seems to work well for the task at hand. I have provided on the handout some Web sites that have more information about link validation software. While I don't want to discourage anyone from exploring the various options available, I do suggest that you investigate what the Webmasters at your own institutions are using. It is likely that they are using some kind of program, and its nice to have some systems support when trying to negotiate software with this level of sophistication.

This is what a typical LinkBot report looks like: <**www.willamette.edu/~aweible/dlc/testpgrpt.htm**>. This report is based on a different file I created to help illustrate how the software validates PURLs. In relation to PURLs, the important aspect of link validation software that needs to be determined is how it handles redirected links.

In the case of LinkBot, PURLs are listed under the heading "Warnings." In this section of the report, we see that the software is alerting us to the fact that the PURL is actually going to the URL. After checking with the company, we were able to determine that the software does then go on to check the validity of the links it is directed to from a PURL. If there is a problem with the URL connected to the PURL, it will appear in the appropriate section of the LinkBot report, usually under Broken URLs. What constitutes a broken link in this report varies, from "source not found" to "server down" to "timed out."

I use the information found in the LinkBot report to follow up on problem links. This requires the sometimes time-consuming process of searching for the new URL for resources that have been moved. I usually approach this task by surfing the agency's Web page, playing with variations in the URL address, and if necessary, sending an e-mail to the agency's Webmaster. Once I locate the correct location, or determine that the resource no longer exists, I then correct the information in the library's catalog record. I also make a special attempt to alert GPO, via askLPS, when I find PURLs that need to be updated.

While this process works relatively well in helping to identify broken links, it should be obvious that link validation software can only go so far in determining whether links in the library's catalog remain accurate. At their best, they can tell you whether there is still a valid file associated with a particular address. They still cannot tell you if the content of the file is the actual resource described in the bibliographic record.

A good illustration of this problem can be seen in one of GPO's old practices related to PURLs. According to my sources at GPO, this is no longer the current practice, but originally, when a link was identified as broken, and no alternate location for the resource could be found, GPO updated the PURL record so the user will be redirected to a Web page that looked like this: <**www.willamette.edu/~aweible/dlc/deadlink.html**>. This "deadlinks" page is a valid link. Since LinkBot can't read the content of this page, it will not recognize the URL as "broken." This was a problem because in my own library, I do not want to direct a user to a page like this. I would prefer to remove a broken link completely from a catalog record, or remove the whole record, if appropriate. But, because LinkBot could not identify this as a problem, I was not alerted to the status of the link, and could not perform my own maintenance to resolve the problem. This is why I had chosen, in most cases, to use the URL, rather than PURL, in our library's catalog records.

It is my understanding that GPO no longer places a link to the deadlinks page in PURL records, but instead places notes on both the Browse Electronic Titles page and the catalog record when a dead link is identified. From the perspective of my own link maintenance activities, I think this is a positive change. I have a bit more confidence that the LinkBot software will identify the broken links within a PURL, and as a result, I will be more likely to leave a PURL in our catalog records, instead of replacing it with the original URL, as had been my previous practice.

I do hope that GPO is able to go back and revise the PURL records that still point to the deadlinks page, so that the records conform with current practices. As we heard in the conference program earlier today, the Electronic Collection Team appears to be working on refining policies and procedures related to the management of the Electronic Collection, and I hope this includes a review of PURL creation and management activities.

This leads me to one final observation about link maintenance activities. Remember when I said earlier that one of the reasons GPO implemented PURL technology was to save the labor costs associated with editing catalog records? Well, I think GPO staff would agree that any labor savings they may have gained is quickly being lost in the truly labor intensive activity of keeping links up to date. Without a completely automated system to accomplish this work, and it doesn't look like there is a magic software solution just around the corner, it is clear that considerable effort needs to be expended to ensure the validity of links to resources in the FDLP Electronic Collection.

Who is actually responsible for this work remains an issue, but I would like to advocate that the depository community has a vested interest in sharing the burden of this work with GPO. It is absolutely necessary that librarians report broken links to GPO as they are discovered, for there is no automated process that will ensure that the links will always be identified.

At the same time, I believe that GPO needs to consider opening up the process of link maintenance, so that libraries with established processes for link verification can contribute more directly to the work that needs to be done. One suggestion would be to add a link to askLPS directly on GPO’s PURL Server pages, to help facilitate the reporting of broken links.

Improvements to the search capabilities of the PURL server, as well as the addition of more data fields like OCLC number or title, would help librarians to use the server more effectively in link maintenance activities. It would also be helpful to have a direct e-mail link to the staff members who work with PURL maintenance, again to help facilitate communication.

Another possibility would be to have GPO authorize particular libraries with the ability to access and edit records in their PURL server, so that records can be directly updated without having to go through the sometimes cumbersome process of reporting information through askLPS. At the very least, GPO needs to share information about current policies and procedures, establish as much consistency as possible, and work to bring old records into compliance with new policies. Libraries must also share what they know about link maintenance, and communicate their needs to GPO.

So, in conclusion, I’d like to summarize the points I have tried to make in my comments about link maintenance activities. There is no question that link maintenance is a labor intensive activity, and it is essential that before these activities are undertaken at the local level each library discuss the importance of accurate links, and evaluate this in terms of the labor costs involved in pursuing such a goal.

An important fact to consider in this discussion is that GPO’s implementation of PURL technology does not guarantee depository libraries that all the links provided in GPO catalog records will remain accurate, and even vendors, at this time, do not provide an absolute solution to the problem. Based on the information Nan, Tom Tyler, and I have collected, it is clear that link maintenance procedures depend on so many local variations in systems and Web support services that it is nearly impossible to recommend anything but the most general suggestions for link validation activities. I hope that the description of my own institution’s procedures are helpful in that regard.

Finally, I want to remind you that all this hard work can pay off. By undertaking link maintenance activities on the local level, and especially if you communicate the results of this work to GPO, you also provide a benefit to the rest of the depository community. Building a partnership with the depository community is probably the only way that GPO can realistically accomplish its goal to keep links to Web resources accurate.

As partners, both sides need to work together to make sure that the policies adopted by GPO work with both GPO and library procedures. More work and communication needs to be done, but given that we are all learning how to make do until technology catches up, I think we can be pretty proud of GPO and the depository community's leadership in trying to find solutions to the issues associated with link maintenance in library catalogs.

I also want to put in a final plug for Tom Tyler’s fine article, "URLs, PURLs & TRULs: Link Maintenance in the Web-accessible Catalog" <**www.du.edu/~ttyler/cil99/proceedings.htm**>. It addresses some issues I’ve just touched on in more detail, and has more evaluative information about particular library catalog systems and link checking software.