Citing and Linking in Electronic Scholarly Publishing: A Pragmatic Approach

Mark Doyle (doyle@aps.org)
The American Physical Society, 1 Research Road, Ridge NY, 11961 USA

The transition from traditional, paper-based publishing to electronic publishing presents many challenges for publishers and their constituents. Among the most significant are those that concern linking to articles that are available online and citing of articles that are now published online prior to (or even in lieu of) the final pagination and binding of paper issues. The solutions to these problems will have a great impact on the usability of the scholarly electronic corpus for the research community. In particular, the solutions must be easy to use, persistent, and scalable. Here we examine some pragmatic solutions (in use now) that satisfy these criteria and contrast them with other proposed solutions. The examples are drawn mainly from the experiences of the American Physical Society, but the lessons gleaned will have wide applicability.

Introduction

By and large, scholarly publishers have achieved their first goal in moving from the "papyrocentric" [Harnad] past to the electronic future. Namely, we have been able to "shovel" [Okerson] electronic copies of the articles we publish in paper form onto the web for delivery directly to researchers' desktops. Now we are faced with the more difficult task of weaving together all of these articles into a seamless web of discourse, easily navigable by the researchers who need to make use of the information. Central to this is the issue of naming articles (or components thereof) and providing ways to robustly link to them together.

Traditionally, scholarly journal articles have identities associated with them as a result of a peer-review process that culminates in its appearance in a bound, paginated paper issue. The article is identified by a natural hierarchical scheme such as (publisher, journal, volume, issue, page). This may have to be supplemented with additional information such as an author's name or a sequence number (position of an article on a page) to get fully unique identities. Once such a designation is given to an article by a publisher, the use of this identity gives researchers a good idea of the quality of the article and where to find it.

Common citation practice is to suppress some of the redundant information such as the publisher, the sequence number and, quite often, the issue, leaving, for example, just the triplet of (journal, volume, page). Bibliographies usually contain just this triplet of information, along with author names and the year of publication. From now on, I will use the idea of this triplet as a concrete model and mnemonic for the more precise idea of a "subset of an article's bibliographic metadata traditionally cited by researchers in a particular field." This subset, of course, can vary from journal to journal and field to field with analogous "triplets" being (journal, volume, issue, page) or (journal, volume, page, author).

This identification scheme and the use of abbreviated citations has been stable over hundreds of years. Electronic publishing presents new challenges for this scheme, however. How does one cite an article which is published online well in advance (or even in lieu of) the pagination and binding of a paper issue? How can the triplet be associated to a robust URL or other pointer so that the document can be retrieved? Can either of these be done in a human friendly way?

The American Physical Society, as publisher of the *Physical Review* journals and *Reviews of Modern Physics*, has chosen to take a pragmatic approach to dealing with the issues of linking to our articles and of separating naming from print production. We do not wholly abandon the traditional triplet that has served us so well - and which researchers still rely so heavily upon. The scheme we have developed has the virtue of simplicity and thus libraries, publishers, and researchers who want to link to our journals have been able to do so with minimal effort. Of course, our solutions don't pretend to provide the most general solutions to the problems of making information available on the web, but in the trade-off between generality and sufficiency, we believe we have struck a happy balance.

Two aspects of our solution are described. The first is the APS link manager which is designed to be a simple and consistent way to link to any article published by the APS (and I do mean any article - we intend to go all the way back to 1893, the beginning of *Physical Review,* in our Physical Review Online Archive project [PROLA]). The link manager is designed to provide simple, persistent URLs to access an article, no matter where the article is actually located. The other aspect is our "e-first" numbering scheme for handling articles that appear electronically before (or in lieu of) paper production. This scheme is used now in *Physical Review D* and *Physical Review Special Topics - Accelerators and Beams* and will be adopted by our other journals as well. The principle design criteria of this scheme is that it fit naturally into the traditional triplet, that it be media independent, and that it can be used to locate items easily in either the electronic or bound paper journals.

Throughout the paper, the point of view will be that of a pragmatic user outside the development process for the various "standardized" solutions. In comparison with some of these other initiatives (discussed below), our approach may appear somewhat pedestrian. But it has the merits of working now and it can be (and has been) readily adopted by others wishing to link to the APS journals.

The APS Link Manager

The APS link manager [APS] was designed to solve one simple problem. To maintain flexibility in developing our electronic journals we had chosen to use a combination of different vendors and in-house resources to put our various journals online. Our vendors use a range of URLs ranging from the simple to the very complicated, the latter requiring knowledge metadata not found in the traditional abbreviated triplet citation.

The vendor URLs have varied with time, undergoing changes as the products evolve and as we move products among our vendors. In order to insulate users from needing to know extraneous information like the issue number, from URL changes, and from what platform a particular journal was deployed upon, a simple URL based upon just the triplet was developed. These URLs are guaranteed to be persistent and robust - if we were to change how a journal was deployed, the linking URL would remain the same but the link manager would be updated to handle the new information. No outside agent would have to adjust their URLs.

For example, to link to the citation Phys. Rev. D 55, 1 (1997) , one would use the URL

http://publish.aps.org/abstract/PRD/v55/p1

Although this URL looks like it might point to a static page, it does not. /abstract/ is mapped to a CGI program that handles the request. The name is descriptive because this is what the user will actually see upon following the link. Our fundamental target for linking is a wrapper page that contains the basic information about the paper, including the abstract. The wrapper can also contain links to any errata or comments, companion papers, references, forward citations, and all deliverables (PDF, HTML, FAX, scanned images for archival material, XML/SGML in the future). The APS has promulgated a policy of making the basic wrapper freely available to all so that these links will work for anyone. Authentication takes place when a user tries to access material that actually requires a subscription such as full-text. The wrapper concept is a natural way to deal with having multiple deliverables and a variety of related information for any particular article. Even more importantly, citation-based linking to a wrapper extends the notion of the citation beyond just being a name for the article itself to a broader concept of a path not only to the article, but to all of the services and other material associated with it.

The rest of the URL is the identifier for the article which maps in the obvious way to the triplet. If one has additional metadata, one can include it in the URL as well. For example, an issue can appear in a /i...../ field and a year can appear in a /y...../ field. These fields can serve as redundant checks on the validity of the citation (offering clues to decoding an occasional invalid citation), but they are not essential to the basic link. The components of the URL after the /abstract/ can be given in any order and they are case insensitive. Leading zeros are also handled gracefully, as well as idiosyncrasies like prepending an 'R' to a page number for a Rapid Communication. There is also an /e...../ field for "e-first" identifiers which will be discussed below.

The chief advantage of the APS linking interface is that the URL is derivable algorithmically from the typical abbreviated citation. No extensive lookup tables are needed by those trying to link to *Physical Review* . This means that anyone can trivially produce an accurate link to any Phys. Rev. article. This feature coupled with the free wrapper concept means that there is no barrier for any individual or agency to link to our journals.

This type of interface is also readily extensible to other deliverables merely by replacing the /abstract/ with something else. For instance, linking to a PDF file could be done through /pdf/ or to a table of contents through /toc/. Our current policy is not to supply a robust interface for full-text deliverables like PDF files. We would rather have users come into the wrapper where we have additional links of interest and give users a pathway into the journal. If a researcher wants to navigate around a bit, she can do so (adding back in the serendipity factor familiar to researchers who discover an adjacent article of more interest than the one she was looking for). This also protects the *Physical Review* "branding" that derives from having the articles gathered together into a single integrated collection with an interface controlled by the APS.

My original model for this type of URL came from my experience working on the Los Alamos e-Print Archive[LANL]. The URL structure there has been stable for over 5 years (an eternity in cyberspace) and has resulted in robust virtual overlays built on top of the archive. A citation of one e-prints by another is trivially turned into a hyperlink allowing for extensive, automated hyperlinking within the archive itself.

Dumb vs. Intelligent identifiers

The link manager made the pragmatic choice of using the triplet as the interface for linking to our journals. It leverages the entrenched citing scheme that is universally used in our field. The syntax was chosen to be transparent and URL-friendly, requiring no special encoding of characters. Over the years, there have appeared alternative identifiers that attempt to uniquely label published items [Paskin, Greene]. A representative selection would include the SICI (Serial Item and Contribution Identifier) [SICI], PII (Publisher Item Identifier) [PII], and DOI (Digital Object Identifier) [DOI] schemes which all define a labeling system for identifying objects.

The goal of the SICI is to provide a uniform syntax for arranging the citation metadata of a serials article into an identifier string. From our practical viewpoint, it has the benefit of being closely tied to the traditional citation of a scholarly article. This makes it an "intelligent" identifier. A full SICI is usually expected to have more than just the triplet information to ensure uniqueness. For instance, the SICI allows for the inclusion of the first letters of the first six words of the title of the article. When the available metadata is limited, as in an abbreviated citation, a SICI can still be formed. This has the consequence that items may have multiple SICIs. The SICI also has some extra information about the format that the information is in, so different deliverables may have different SICIs. Limited SICIs could serve as the basis for a link manager (in fact, it would be straighforward to provide a /sici/ interface), but there seems little advantage to adding to the complexity of the URL syntax. Indeed, the SICI has the added problem that URLs using it need to have the special characters incorporated into its syntax (i.e., the characters "<", ">", ";", and ":") encoded for proper use [RFC]. No agency has ever approached us to implement a SICI -based interface.

The PII tries to provide a unique identifier for an item without the constraint that the item be given a final traditional citation. Thus, a PII can be assigned to an article before it is published. However, this goal lead to PIIs that don't have any intelligence built into them. The APS was part of the collaboration that developed PIIs and every article we print has a PII printed at the end of its abstract. However, we have made no use of them and we don't even track them in our manuscript database. Since PIIs are completely unrelated to the triplet, using them for linking would necessitate an additional step where the triplet is mapped to the PII. Agencies wanting to link via PII would then have to maintain a database of PII lookups (if they did it dynamically, there is no advantage because the interface for retrieving a PII would have to have the same properties as the present link manager, greatly diminishing the usefulness of the PII).

The DOI initiative is much broader in scope than the SICI and the PII initiatives. Reaching beyond just naming objects, this initiative also tries to bring in a richer set of metadata. In particular, DOIs are meant to serve as a way of conveying information about the ownership rights (e.g., copyright) associated with an item as well as providing a way to locate it. The developers of the DOI have wrestled quite a bit with issues that are especially difficult when a general solution is desired. There is a great deal of debate about the tradeoffs between putting intelligence into the identifier versus putting it into the metadata and about deciding how to associate a DOI to collection of deliverables and derivative products [Bide, Paskin].

DOI identifiers have two major parts. The first part gives some information about who originally produced the DOI. The second, more important part for this discussion, is a string of characters that serves as the identifier. Many have argued that this part of the identifier should have no intelligence at all. Others have decided that intelligence is needed for practical reasons and have resorted to using, for example, the SICI as the second part of the DOI. One could also create DOIs based upon link manager style URLs. Both of these questions raise the problem of which metadata should be included in the DOI itself and whether that metadata is widely used in citations to allow easy linking.

The other distinguishing feature of the DOI initiative is that it has an initial implementation based upon CNRI's Handle System [CNRI]. Given a DOI, one can trivially construct a URL that will contact a centralized server and database and map the DOI to a publisher's URL for the item. Thus, DOIs can directly serve as linking mechanism. The present implementation maps a DOI to a single URL, but this doesn't have to be the case [Paskin2]. Some publishers are already using DOIs for linking to articles that have been published online, but not yet paginated. Some of the drawbacks to this approach will be discussed below.

Indirection

The central feature which is common to all reasonable linking schemes is indirection. That is, the link goes first to a resolving engine which furnishes the final URL for the

item and then redirects the user using standard web technology. Any linking scheme which is to provide robustness and persistence must incorporate indirection. Services which promote linking without indirection will likely face future problems with broken links or with an inflexible, legacy URL scheme that can't be changed. Thus, it is quite natural that the APS link manager and the DOI incorporate indirection in much the same way.

OCLC has developed Persistent URLs (PURLs) [PURL] which are similarly based on indirection. PURLs, like DOIs, are implemented through a set of centralized servers which map a PURL to a real URL. PURLs, however, are neutral about how items are named and how to associate PURLs with the names. The link manager scheme would fit naturally into the PURL scheme, though there seems to be very limited advantages for a publisher to go through the PURL mechanism rather than just providing its own persistent scheme. One slight advantage is that if a publisher goes out of business, the PURLs will still be available, but instead of resolving to a URL, they would return a history of the URLs that had been associated with the PURL.

This marginal benefit needs to be weighed against the gains the publisher has in tying their link resolver directly to their definitive manuscript database. Because of indirection and the localization of the resolver at the publisher, the APS link manager is given the chance to resolve errors. In particular, when an improper citation is given, the user is given an informative error message (e.g., a choice of likely corrections) so that they don't just hit the '404 - Document Not Found' brick wall. In addition, the user at least finds themselves at a site where they can begin to look for the item they are after, perhaps by following links to a search engine. A publisher could also maintain a list of its mirror sites and offer the user the choice to redirect to a convenient one, or even let frequent users register and configure which mirror site they want.

This brings us to the crux of the matter. Indirection is an essential feature of robust linking systems and so all of them incorporate it. The aspects that demand careful consideration and evoke debate are the interface to the indirection (triplet, full metadata (intelligent or compound identifier), or no metadata (dumb or simple)) and the location of the resolvers (centralized, as in DOI and PURL, or distributed at each publisher as in the APS link manager). The choices made here have a great impact upon the ease and scalability of linking. A strong case can be made for using intelligent identifiers based on the triplet and URLs associated with a publisher's own resolver.

If a publisher's linking interface requires either a dumb identifier or full metadata, then those wishing to link to the publisher will require a way to map triplets to the appropriate identifier adding an additional layer of indirection. This will require the publisher either to provide a triplet-based interface to give out the identifiers, to develop some other method of supplying the data on a one by one basis to potential linkers, or to contribute to some centralized database that does the mapping (with an interface based on the triplet). If one is going to provide a triplet based interface anyway, it might as well include the linking information obviating the need for using the

identifier as an extra level of indirection for linking. Supplying linking data to all potential linkers means that everyone has to maintain a database of your identifiers; this is resource intensive and not very scalable.

Contributing to a centralized database means that you have to maintain your data and ensure its integrity. Furthermore, you give up flexibility in how you resolve queries based upon erroneous information or to intercede with options about mirroring or other preferences. The current implementation of DOIs have the further drawback that you have to supply an entry for every one of your items rather than just supplying an algorithm based upon intelligent identifiers. PURLs seem to have similar drawbacks, but they do allow for partial redirection. In other words, only part of the PURL is handled on the centralized server - the rest is passed along as additional information to whatever URL is associated with the PURL. In that case, the advantages of having a PURL in the first place are minimized-- why not just go directly to your resolver?

Because the APS interface is built upon the triplet, there is no need for an agency linking to us to maintain information about the link itself. They need only have a description (algorithm) for transforming the triplet to our URLs. Linking can become quite dynamic, done on the fly if necessary. If centralized databases become commonplace (it seems to be the trend at the moment), then we have the strong requirement that we would want to be able to supply just our algorithm for creating links based upon our triplets. We don't want to have to maintain hundreds of thousands of identifiers and URLs on remote servers.

E-first Identifiers

One reason publishers turn to identifiers like PIIs or DOIs is that they would like to provide a citable reference to an article so that it can appear online before it is paginated and bound into the paper journal. Publishers who choose this path end up in one of two situations: their articles are cited in two different ways depending on whether it was used before or after paper publication or they abandon the pagination altogether and perhaps end up with cryptic identifiers (some quite lengthy) that are unnatural (even abhorrent) to researchers. For instance, Springer-Verlag has chosen to do this with The European Physical Journal C. An example DOI URL for such a paper would be http://dx.doi.org/10.1007/s100529901100. When this paper appears in a paginated, bound issue, it will get a second, more traditional citation with a page number.

The American Physical Society chose a different path. We wanted an identifier that would be media independent, separate from the printing of a paper journal. This would enable us to effortlessly drop print if that becomes desirable in the future. Furthermore, there should only be a single, definitive way to cite the article and, because we are in a transition period where print still plays an important role, an article cited by its electronic identifier should be easy to locate on a library's bookshelf. The success of the identifier would also be tied to how comfortable researchers would be in using it directly in their citations. Finally, we wanted to be able to preserve some of the features

of a print journal such as grouping articles in tables of contents by subject, even when the articles are listed in identifier order for easy retrieval.

The solution, conceived by Erick Weinberg, an editor of *Physical Review D*, was to come up with a simple six digit number that can be treated as if it had no structure, but in fact there is internal meaning that is used for ordering articles in an issue. A typical number would '034013'. The first pair of digits represent the traditional issue, the second are a subject classification, and the final pair is the article number - it is just a sequence number within the subject for that issue. Thus, a Phys. Rev. D citation in the new scheme would be Phys. Rev. D, 58 034013 (1998) with the associated URL http://publish.aps.org/PRD/v58/e034013. When (and if) articles are printed, they are bound in sorted order and the articles are paginated internally as 034013-1, 034013-2, etc. The link manager than uses /e034013/ for this article in lieu of the page number (in fact, since the identifiers look like page numbers, the link manager is quite tolerant of whether /e..../ or /p...../ is used and the presence or absence of the leading zero).

Since introducing this scheme in July 1998, the APS has received very positive feedback about this pragmatic solution. It preserves the qualities of traditional citations which has made them stable over hundreds of years and it fits in well with what researchers know. Furthermore, third-party abstracting, indexing, and citing services and libraries can easily accommodate it, even though it breaks the print paradigm.

Other publishers and users

From its introduction in September 1997, the APS link manager also has been enthusiastically accepted by the library community. SLAC's SPIRES database [SPIRES] immediately began using it to link to any Phys. Rev. articles in its database. Datastarweb and the CERN library incorporated APS links into their offerings [CERN] almost immediately. The library has established an automatic process for extracting references from the PostScript versions of papers they catalog. By using the link manager, they have been able to quickly turn thousands of citations into hyperlinks to APS papers. Because of the minimal effort it takes to incorporate our links into their services, Jens Vigen and his CERN library colleagues have become quite vocal in lobbying other publishers to adopt a similar interface to their journals. The University of California Melvyl catalog has also made use of the link manager to tie their service to our journals. Individual researchers are now linking their curriculum vitaes directly to the journal articles.

Other publishers have also begun to create their own link managers. Academic Press has a very similar link manager in place and we have collaborated with EDP Sciences, Springer-Verlag, and IoP Publishing on their linking projects. The effort involved in creating a link manager is directly related to the type of information used in URLs a publisher uses in their online journal products and whether they have in place a database that can map traditional citation information to the information in the URL if it extends beyond the traditional information. If the requisite database already exists,

then the a working prototype link manager can be developed within a week.

Having the link manager in place has greatly simplified the APS effort when it comes to mutual agreements for inter-publisher linking. There is no obstacle to a publisher simply reading our description of the link manager and implementing it. Typically we ask that publishers create a similar interface so that we can link to them. Needing to maintain an extensive database or lookup table of another publisher's material represents a large obstacle from our point of view. Such solutions don't scale very well - if there is to be interlinking between many publishers, each would have to maintain a database for every other publisher (or, possibly, publishers would group together and contribute to a shared database which each publisher would to maintain).

The link manager concept also opens up further possibilities for new offerings and services built around our journals. One can envision virtual journals or collections of common papers (e.g., all Nobel prize papers) which combine articles from different journals or from different time periods into a single offering. By incorporating indirection, one can easily provide alternative access restrictions for those coming through a different portal. Another idea is to allow libraries and other organizers of information to index our material (full text SGML for instance) so that their users can use familiar search engine to search across multiple publishers. Retrieving a search result would then take place via the link manager to the article on the APS service.

SLinkS

With publishers choosing a variety of linking schemes, there is a need for a way of sharing information about how a particular publisher's link manager operates. Eric Hellman of Openly Informatics, who will be giving a talk in this same session, has developed a metadata description for describing link managers of all types call "SLinkS" [SLinkS]. Based upon the W3C's recent RDF [RDF] recommendation, SLinkS provides for URL templates which can be filled in with appropriate information to achieve the linking. However, his specification takes an agnostic point of view about what information should go into building the URLs and where it should come from. Lookup tables, for example, are accommodated by the scheme. While the SLinkS proposal has the real potential to allow publishers to extensively link articles, this potential can really only be fulfilled if publishers adopt URL templates that depend only upon the information that appears in traditional citations and that generate persistent URLs. Thus, one can envision a world where if a publisher would like to provide links to another publisher's journals, one can just query the publishers' SLinkS description and plug in the citation information and generate a link. Such a state of affairs would quickly allow publishers to hyperlink the vast majority of scholarly articles into a single web of discourse with great benefit to the research communities which we serve.

Conclusions

Although quite simple in idea and implementation, it should be clear that there are

many compelling advantages to taking a pragmatic approach to providing a link interface to scholarly journal articles. The key concept of indirection coupled to a publisher-based resolution, enabling validity checks using a definitive database, provides robustness and persistence. Localizing at the publisher level and providing interfaces based upon traditional citation information improves scalability and greatly reduces implementation time.

Electronic identifiers that extend, but do not stray far from traditional citing strategies, even while breaking the print paradigm, can be quickly adopted and maintain simplicity. This, of course, is not to say that these pragmatic approaches solve or address all of the same issues as the alternative schemes. However, the practicality of applying generalized solutions to limited domains like the citation and linking of scholarly articles remains dubious, especially if they present barriers to rapid adoption. These alternative identifiers were developed with a wide range of goals in mind and all of them have shortcomings from the pragmatic point of view of providing a simple, robust basis for linking to peer-reviewed scholarly articles, a rather narrow domain, with a well-developed mechanism for identifying articles. Their shortcomings derive mostly from their generality and the subsequent need to incorporate another layer of indirection. The examples given embed, by design, lessening degrees of intelligence in the identifier. These require a fair amount of additional infrastructure and will permanently require dual citation mechanisms since researchers will be reluctant to use dumb identifiers in their every day work.

Even if dumb DOIs become ubiquitous, it seems that it will always be necessary to be able to map citeable subsets of metadata to such identifiers. Furthermore, the URLs that a DOI resolves to should be as persistent and robust as possible to reduce the need to maintain the data on the DOI servers. For both tasks, a link manager style interface will be invaluable and, thus, publishers have little reason not to take this simple step towards creating the ultimate web of scholarly discourse.

References

[APS] *Frequently Asked Questions about the APS link manager* (http://publish.aps.org/linkfaq.html)

[Bide] M. Bide, "In Search of the Unicorn: The Digital Object Identifier from a User Perspective," *British National Bibliography Research Fund Report, Book Industry Communication*, London, November 1997. (http://www.bic.org.uk/bic/unicorn2.pdf)

[CERN] *CERN Library Catalog* (http://weblib.cern.ch/)

[CNRI] *CNRI Handle System Resolver* (http://www.handle.net/resolver/index.html)

[DOI] *The Digital Object Identifier* (http://www.doi.org/)

[Greene] B. Greene and M. Bide, *Unique Identifiers: a brief introduction*. (http://www.bic.org.uk/bic/uniquid)

[Harnad] Coined by Stevan Harnad, see for example "American Physical Society's

Discipline-Specific Model for Electronic Publishing," *Serials Review, Special Issue on Economics Models for Electronic Publishing,* pp. 58 - 61 (1992). (http://cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad92.interactivpub.html)

[LANL] *Los Alamos e-Print Archive* (http://xxx.lanl.gov/)

[Okerson] Ann Okerson, *Recent Trends in Scholarly Electronic Publishing,* Yale University (1997). (http://www.library.yale.edu/~okerson/recent-trends.html)

[Paskin] N. Paskin, "Information Identifiers," *Learned Publishing,* vol. 10, no. 2, pp. 135-156 (April 1997). http://www.elsevier.com/inca/homepage/about/infoident/)

[Paskin2] N. Paskin, *DOI White Paper - Current View and Forward,* DOI Discussion Paper, version 3 (August 1998). (http://www.doi.org/white-paper-3.pdf)

[PII] *Publisher Item Identifier* (http://www.elsevier.nl/homepage/about/pii)

[PROLA] *The Physical Review Online Archive* (http://prola.aps.org/)

[PURL] *Persistent URL Home Page* (http://www.purl.oclc.org/)

[RDF] *W3C Resource Description Framework* (http://www.w3.org/RDF/)

[RFC] RFC 1808 (http://www.w3.org/Addressing/rfc1808.txt)

[SICI] *Serials Item and Contribution Identifier Standard* (http://sunsite.berkeley.edu/SICI)

[SLinkS] Eric Hellman, *ScholarlyLink Specification Framework (SLinkS)* (http://www.openly.com/SLinkS/)

[SPIRES] *SPIRES Library Databases* (http://www-spires.slac.stanford.edu/find/spires.html)