Learned Publishing (2005)**18**, 177–185

The transformation of scholarly communication

Simeon Warner
Cornell Information Science

ABSTRACT: Recent debate on the reform of scholarly communication has focused on access issues. Although important, access is only one dimension in which the scholarly process can be transformed. Scholars are embracing highly collaborative and data-intensive standards of practice influenced by powerful computing and network technologies. This dramatic transformation of scholarship demands a natively digital, network-based scholarly communication system that is able to capture the scholarly record, make it accessible, and preserve it over time. I will offer a technological perspective on how these demands might be met.

A note about access

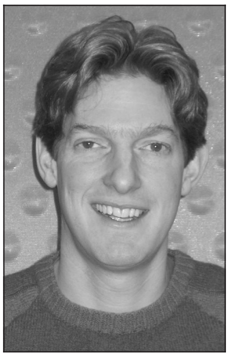
We are all aware of the current debate about open access to scholarly material and about business models that might facilitate and sustain this. Although open access is not the focus of this article, I think there are a couple of pertinent observations to be made:

- An increasing number of subscription access journals are making data available to search engines such as Google. This is a demonstration of the understood value of data availability for the provision of services such as global searching. Such data sharing can be permitted without sacrificing established business models.
- There is now strong evidence that openly accessible articles have greater impact (usually measured by citations[1–3]). The normal way to interpret this it that open access results in more readers and hence greater impact, although one might also argue for a self-selection effect in that the better researchers are the ones who make their material available openly. It is not clear what fraction of the extra citations come from readers who have access only to open access material, and what fraction are because the work was just *easier* to find or access.

A stovepipe view of scholarly publishing

Consider the value chain of functions fulfilled by publishing an article in a journal shown in Figure 1. The author submits an electronic version of an article to the *Journal of X*, the uploaded article is stored and the time of receipt recorded. The article is then sent out for peer-review and there may be one or more iterations of correction and review. If accepted, the journal staff then arrange copyediting, preparation and formatting. Finally, the formatted article is made available to subscribers electronically.

Based on a keynote presentation at the ALPSP 21st International Learned Journals Seminar, The Web and after: the future of scholarly e-publishing, London, 8 Apr 2005, http://www.alpsp.org/events/biogs080405.htm#war



Simeon Warner

RIVER VALLEY ADVERT

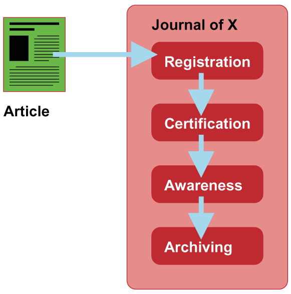


Figure 1. A simplistic view of scholarly publishing showing a value chain of functions.

Paper copies may also be printed and mailed. In the electronic-only case the publisher keeps an archive and in the print case the publisher and many libraries preserve copies in different locations. However simplistic the picture, we see that this arrangement neatly packages the four functions required for scholarly communication as identified by Roosendaal and Guerts.[4]

The stovepipe with bells and whistles

Figure 1 is of course far too simple. It does not take into account any of the other services typically associated with a scholarly journal. Figure 2 adds several other services. It adds archiving provided by a national library, an increasingly common solution to address fears of electronic resources being lost if the originating organization were to vanish. It adds abstracting and indexing services to provide greater awareness and access, often allowing inclusion in library search services via federated search. A review journal provides a second level of certification and an additional awareness pathway. In mathematics, for example, *Mathematical Reviews* and *Zentralblatt MATH* provide important secondary certification and awareness.

A fifth function, *rewarding*, is also shown in Figure 2. Journal recognition and prestige obviously have a significant impact on the impressions of tenure and hiring committees. However, the ISI *impact factor* is an (over-)popular method used to provide a seemingly objective measure of the significance of journals which is then used to estimate the impact of articles, based on where they are published, for academic rewarding. Calculation of the impact factor relies upon access to citation data for many journals.

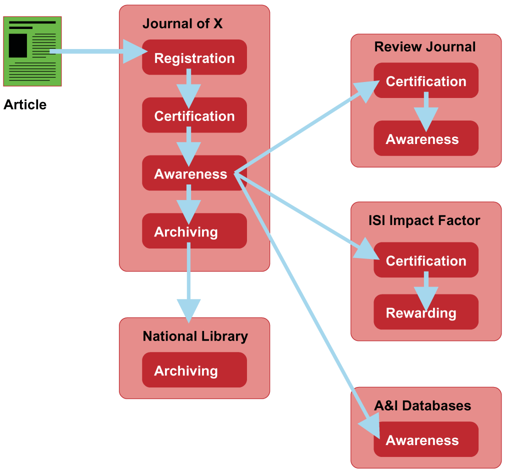


Figure 2. An extended view of the scholarly publishing value chain where additional services facilitate additional pathways.

Interoperability now: of PDF and DOI

The current state of interoperability is limited. There has been considerable adoption of identifier systems (e.g. DOI) and of standard electronic formats (e.g. PDF). The use of PDF means that users need only one viewing technology for all journals. A positive side-effect is that services that can index PDF manage to provide quite effective search tools based on harvested content, even if the PDF is semantically much poorer than most source formats.

Digital Object Identifiers (DOIs) provide an acceptable way for competitors to provide persistent links between different systems via a fair intermediary. DOIs also provide a new identifier space (separate from URLs) within which scholarly works may be identified. Google Scholar uses DOI-based links to link and to index articles from the American Physical Society (APS) and other CrossRef members. By providing DOIs for cited articles, APS articles are appropriately ranked using Google's PageRank[5] algorithm applied to DOI-space rather than URL-space. Unfortunately, web clients do not natively understand DOIs so services must translate them to URLs using a particular

journal recognition and prestige obviously have a significant impact on the impressions of tenure and hiring committees service (a resolver). Thus many different URLs may exist to resolve the same DOI. For example, links to APS articles in Google Scholar explicitly use the APS resolver rather than the default DOI federation resolver.*

Further interoperability is provided by federated search (using Z39.50[6] and newer forms under the ZING[7] umbrella) or by metadata harvesting (using OAI-PMH[8]). Use of these technologies has been impeded, to the detriment of users, by financial models based on selling metadata, the desire to draw users through local portals and perhaps by happiness with the status quo.

OpenURL is a promising technology that has just been adopted as a NISO standard.[9] OpenURLs allow users to be provided with services appropriate to their context, perhaps taking into account identity, permissions and location. We have yet to see how widely OpenURLs will be adopted.

we have yet to see how widely OpenURLs will be adopted

It seems a trivial example to focus on, but I see it as a major failing that I cannot yet go to my local library portal and easily search for a journal article in all their electronic holdings. This is in spite of an expensive federated search engine that searches some (but not all) of 34 subscribed databases. Even search, perhaps the most basic discovery service, is not well supported in a global and interoperable sense. Compare this with the web, where Google allows one to search 8 billion web pages.

Information overload

The phrase *information overload* is used in numerous arguments associated with the selection of appropriate resources from the ever-increasing quantity available. This problem is not a recent concern:

> It is certainly impossible for any person who wishes to devote a portion of his time to chemical experiment, to read all the books and papers that are published in connection with his pursuit; their number is immense, and the labour of winnowing out the few experimental and theoretical truths which in many of them are embarrassed by a very large proportion of uninteresting matter, of imagination, and of error, is such that most persons who try the experiment are quickly induced to make a selection in their reading, and thus inadvertently, at times, pass by what is really good. (Michael Faraday[10])

In some disciplines it may be possible to keep up to date by browsing just a couple of key journals. However, in many, if not most, disciplines such a simple strategy will miss too much. The problem of selection is exacerbated in new fields and in cross-disciplinary research where there may not be established databases or review journals to serve as starting points.

I argue that information overload, the fact that there are too many articles for researchers to sift through, is a problem that is best tackled with the help of automated tools and agents, starting with simple search but including much more advanced techniques. Furthermore, the better the information available – metadata, full text, citation data, certification information, etc. – the better that our agents will be able to help with this selection. The selection problem is strong motivation for the creation of a machine traversable and understandable network of scholarly information.

Changing practices

The practice of scholarship is changing rapidly as researchers take advantage of continuing improvements in computing, storage, networking and data capture facilities. Scholarly communication practices are also changing as a result:

> The traditional, linear, batch processing approach is changing to a process of continuous refinement as scholars write, review, annotate, and revise in near-real time using the Internet.[11]

Such changes demand improvements to the scholarly communication system that will capture this digital scholarly record, make it accessible and preserve it over time. To mirror practices this communication system

*To take a concrete example, consider the DOI 10.1103/PhysRevD.30.272. Google Scholar links to http://link.aps.org/doi/10.1103/PhysRevD.30.272 but one could also use the DOI federation resolver giving the URL http://dx.doi.org/10.1103/PhysRevD.30.272. Both resolve to http://prola.aps.org/abstract/PRD/v30/i2/p272_1.

must change to be more closely coupled to the process of scholarship. It must be able to include rich media, datasets, software and informal documents alongside formal peer-reviewed documents. Flexible access control will be necessary to facilitate easy transition from collaborative work to publication. It must achieve all this without compromising the quality, accountability, rewarding measures and preservation features of the present system. Even the notion of what to preserve becomes complex when we think of continuous refinement of scholarly work. Can the complete evolution be preserved? If it can't, or shouldn't, then how do we decide what should and should not be preserved?

Datasets and 'The Grid'

Data and computer-intensive disciplines are undergoing particularly rapid change. The quantities of data now being used in genomics, high-energy physics, astronomy and climate modelling are staggering. Terabytes are certainly 'last decade', petabytes (10^{15} bytes, a million gigabytes) are in, and exabytes loom large on the horizon (10^{18} bytes: the Large Hadron Collider at CERN is expected to use 1 exabyte by 2012[12]). In parallel with huge quantities of data are huge collaborations which place demands on communication and access infrastructures.

'The Grid' is an set of technologies that provide for large-scale distributed data storage and computation. They are designed to allow seamless data access, migration and caching, combined with flexible and dynamic use of distributed computational resources. At present grids in different disciplines exist separately from the rest of the scholarly communication infrastructure and there is a need to provide interoperability so that data, code and visualizations can be effectively included in the scholarly record.

Recording scholarship

To record scholarship effectively we must extend our notion of the unit of communication from a journal article to include rich media, datasets and software. Bundling rich media (e.g. a video) as an add-on to a publication is clearly not adequate. What is needed is a more flexible approach that permits the composition of complex documents that aggregate and extend other complex documents.

Consider the example of an article presenting an analysis of several terabytes of data stored by the US National Virtual Observatory. A complete record of this work should include not only the article but also the software and dataset. Given the huge size of the dataset it is clearly not feasible to store a copy with the article on the journal site, instead it must be clearly and unambiguously included by reference, in a way that is seamless to the reader.

Key to implementing such a system is the need to facilitate early registration of communication units in manners that accord with community practices. Early registration is necessary to allow seamless transition from collaborative work to broader communication while enforcing appropriate access regimes. Furthermore, the provenance of components of these complex documents (including, for example, formally reviewed components) must be recorded, and the integrity of each component must be ensured and verifiable.

As we expand the notion of what comprises a unit of communication, so must we expand the notion of the preservable unit of communication. In the current system, preservation is typically afforded only to formal documents, mainly peer-reviewed journal articles. Preservation of data and other resources usually happens in local and *ad hoc* ways. In a system that provides for the early registration of communication units we have the possibility of preservation separate from the conventional sequence of scholarly publication. Preservation can be just another service, with different associated pathways and policies. Current work on establishing e-print repositories, institutional repositories and discipline-based data repositories is starting to investigate preservation issues outside of the traditional journal domain.

preservation of data and other resources usually happens in local and ad hoc *ways*

What can we learn from the web?

Two key architectural elements of the web are a uniform linking strategy (familiar http: URLs) and the prevalence of a few standard formats (notably HTML for web content containing links, and PDF and image formats for end-point documents). Add simple-to-use browsers and we have much of the web we know today.

The next step was the addition of robots and spiders that could crawl the web to build indexes for the search engines we now rely upon. It is interesting to note that many times people predicted a failure in the scaling of services to whole-web dimensions, yet web search engines now index billions of documents with no sign of a scaling limit.

Web search engines are wonderful and, much to the horror of some, have become the first recourse for many researchers looking for scholarly works. For certain types of known-item search they are extremely convenient though in others they are dreadful. What determines success or failure?

web search engines now index billions of documents with no sign of a scaling limit

Consider a conversation where your colleague mentions a article by Morgado and Mucciolo about granular gasses. Back at your computer you enter these two names into a search engine. *Voilà!* The first hit is the Eduardo Mucciolo publication list and a few items down you find a direct link to the arXiv article cond-mat/0204084.

Later, you have a similar experience with the mention of a article by Milne and Wells about the light echo of a supernova. Again, you enter the two names into a search engine. This time the results are links pages about *Winnie the Pooh* and oil drilling, nothing about astrophysics in the first couple of hundred hits. Certainly no mention of the arXiv article astro-ph/0210236. What the search actually looked for was documents containing the strings *milne* and *wells*, when we actually wanted only scholarly articles with *author Milne* and *author Wells*. These semantics were missing from the query.

The semantic web

Put simply, the semantic web is about creating a machine-understandable version of the web.[13] This is achieved by structuring information in the form of assertions. A simple example would be *this resource has an author who's name is 'Fred Bloggs'*. The term *has an author whose name is* is identified using a web identifier (a Uniform Resource Identifier or URI) and is actually likely to be rather more strictly defined than suggested above. It may be *has an author, in the sense defined by British Library standard XYZ, whose name is*. It may also be that the author is identified with a web identifier (just as a document is identified with a DOI) and the name, affiliation, etc., are then properties associated with that identifier. In some sense the semantic web extends the idea of metadata to the whole web in the multi-layered and flexible way.

Next, we add to the web of assertions, a set of thesauri and ontologies. A trivial application of this extra information allows the equivalent of the translation of metadata from one format to another (cross-walking): a rule might say that *author in the sense defined by British Library standard XYZ* is the same as *author in the sense defined by the Library of Congress*, where each of these concepts are uniquely identified in vocabularies controlled by the originating organizations (expressed using XML *namespaces*). Thus some agent can understand this equivalence and use the data accordingly. More generally, ontologies provide formal expressions of conceptual schemes which will allow machines to use data effectively. Both the web of assertions and the ontologies are described in the Resource Description Framework (RDF) and this is usually written in an XML syntax.

The next step is to add services to the picture. There has been considerable progress with several established and evolving standards (e.g. WSDL and OWL-S) that aim to facilitate machine discovery and interoperation of services. To interoperate with the semantic web, data must be created with semantic markup and this must be preserved through service pathways. The framework then provides for automatic accretion of semantically rich provenance information for use by downstream services.

Service pathways within arXiv

Why look at arXiv[14] e-print archive? First, arXiv submission, and hence the *registration* of a work, occurs at a time that is flexibly related to registration in the conventional publishing system, and scholars can choose

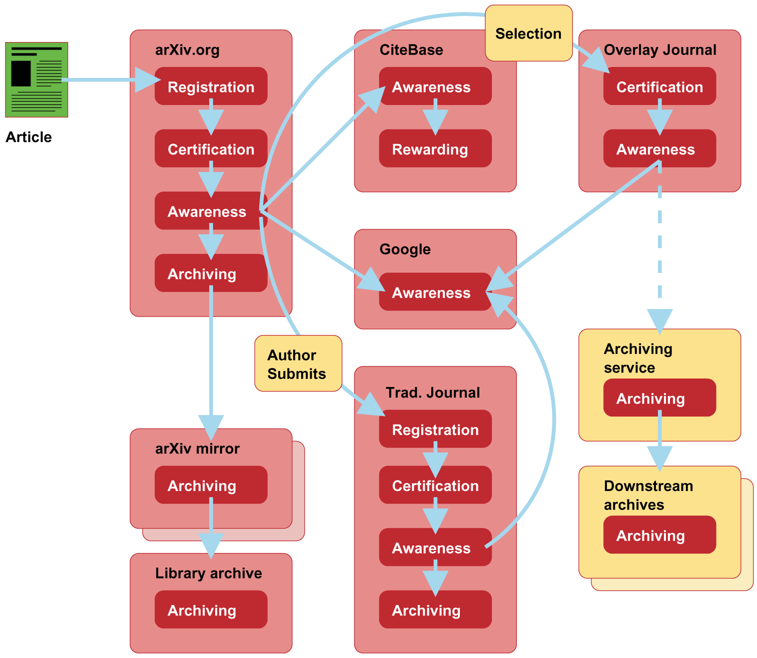


Figure 3. A selection of service pathways based around the arXiv e-print archive. (Adapted from Van de Sompel *et al.*[15])

this relation. Second, arXiv provides open access and has a number of data-sharing services and arrangements which facilitate interesting pathways.

arXiv itself provides most of the functions of the scholarly communication process. It provides registration by recording and displaying the date and time the document was uploaded and then not allowing further undocumented changes. A very weak form of certification is provided by moderation which aims to ensure that *Material submitted to arXiv is . . . of interest, relevance, and value.* This is clearly not conventional peer review but it has proved a useful minimum standard, a certification of sorts. Awareness is provided through then main site, the mirror network and through email alerts. Finally, archiving is provided by maintenance of the main site as part of the Cornell library, through backups (bit preservation) and format migration in the future. The four functions are shown within the arXiv.org box in Figure 3.

Figure 3 also shows a selection of arrows between the different services providing communication functions. The next two sections highlight service pathways fulfilling archiving and rewarding functions.

Archiving pathways

arXiv's archiving strategy has relied upon the combination of adequate redundancy through the operation of a network of separately controlled mirrors, and maintaining online accessibility as technology develops (e.g. the addition of TeX to PDF conversion as PDF became a popular format). We have recently seen the addition of another archiving pathway in that the Bibliothéque Nationale in France is now archiving the contents of arXiv taken from the French mirror site (pathway: arXiv.org to arXiv mirror to Library archive). Furthermore, if articles from arXiv enter the formal scholarly publication system through either traditional or overlay journals, we see additional pathways to archiving.

One can imagine that both arXiv and frameworks such as LoCKSS will eventually rely on the services of additional hubs for the fulfilment of tasks such as digital format migration, which will be an essential part of the long-term archiving function in the

arXiv itself provides most of the functions of the scholarly communication process digital realm (pathway shown from overlay journal to separate archiving service and then to downstream archives).

Rewarding pathways

Figure 3 shows pathways to traditional and overlay journals that would feed into traditional rewarding measures such as the impact factor (not shown). A novel and experimental pathway to fulfil the rewarding function is provided by the CiteBase[16] service which harvests all of arXiv's content, attempts to identify citations through automatic reference extraction and also counts downloads as an estimate of readership. These two metrics have the potential for use in new rewarding metrics. There are valid concerns about the implementation of both of these measures though the ideas are clearly sound (or at least as sound as the impact factor). Automatic reference extraction has become quite robust and one can imagine that this will only improve with better heuristics, greater use of identifiers, and better authoring tools. The situation would, of course, be much better if a truly semantic markup were used in document preparation – in the world of the semantic web we would not have a string that just looks like a bibliographic reference, the elements of the string would be associated with assertions saying exactly what they are: the document identifier, the certification authority, etc. The biggest obstacle to accurate citation counts is that services need access to all the citations for an article to avoid uneven under-counting.

the scholarly communication system must adapt with developing research methodologies

Apart from the obvious objection that downloads are not directly linked to the number of times a article is read, there are serious privacy and reliability concerns about counting downloads. While these statistics are a research topic one can fairly safely count downloads as a measure of real user downloads (with care to remove robot accesses, etc.). If these numbers start to be important, then there is incentive to game the system to artificially inflate them. The crucial difference between downloads and citations is that citations are public (everyone can see the X has cited Y) while downloads are not. Many technological solutions might be applied to measure readership more accurately than by counting downloads though there are serious privacy and openness concerns. Is it right for a client program to 'phone home' with the identity of a user every time a document is opened?

Concluding remarks

The scholarly communication system must adapt with developing research methodologies. It must support collaborative, network-based and data-intensive practices. First, the scholarly communication system must be innately digital. This is already happening for other convenience and efficiency reasons. Second, it must support a much expanded unit of communication that may be heterogeneous and distributed. Thirdly, it must provide for many different pathways that fulfil some or all of the necessary communication functions. The challenge for publishers is to identify functions that really add value, and to implement them in a truly networked fashion that will best serve the community. Added value cannot be claimed in processes that can be automated.

In the near-term, the publication process must become truly digital: semantically marked-up electronic complex documents must be the masters, with display and print versions secondary. Metadata should fall out naturally from such documents and thus need not be seen as a high-cost and hence high-value item, instead it should be shared to facilitate the development of tools and services to better locate and select information.

It is reasonable that scholarly societies help set standards within their disciplines and it is likely that peer review will be a key feature of the landscape for the foreseeable future. However, peer review is expensive and provides only a rough measure of quality (OK or not-OK) that will be used in conjunction with other metrics. We already see other forms of publication becoming important in many disciplines, both because of frustrations with the traditional system and because rich media, software and datasets are better, or can only be, dealt with in other ways. The result is a greying of the line between formal and informal communication. Scholars need a communication system that helps them efficiently publish, access, reuse and assess the relevance and quality of information. These goals will be best achieved by an interoperable system that integrates peer-reviewed publication with other components the scholarly process.

Acknowledgements and disclaimer

This article is based on work done in collaboration with Carl Lagoze (Cornell), Sandy Payette (Cornell), Herbert van de Sompel (LANL) and John Erickson (HP Labs), and on our article Rethinking Scholarly Communication: Building the System that Scholars Deserve.[15] However, for the particular slant given here, only I am guilty. I also thank Mark Doyle for comments.

I acknowledge that my experience and the examples given here are biased toward physics, computer science and mathematics. While the specifics of different disciplines vary, I believe that many of the ideas carry over even if adoption rates will vary widely. Perhaps this vantage point has the benefit of allowing me to see a little further ahead than those in other disciplines. In the particular area of physics in which I last worked, every new article was available from arXiv[14] and thus available to all, usually without delay.

References

1. Harnad S. and Brody, T. Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 2004(Jun), 10. URL: http://www/dlib.org/dlib/june04/harnad/06harnad.html.
2. Lawrence, S. Online or invisible? *Nature*, 2001(411), 521. URL: http://www.neci.nec.com/~lawrence/papers/online-nature01/.
3. Metcalfe, T.S. The rise and citation impact of astro-ph in major journals. arXiv e-print archive, 2005. URL: http://arXiv.org/abs/astro-ph/0503519.
4. Roosendaal, H.E. and Guerts, P.A.Th.M. Forces and functions in scientific communication: an analysis of their interplay. *Proceedings of CRISP 97* (Cooperative Research Information Systems in Physics), 1998. URL: http://www.physik.uni-oldenburg.de/conferences/CR ISP97/roosendaal.html.
5. Page, L., Brin, S., Motwani, R. and Winograd, T. The PageRank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. URL: http://newdbpubs.stanford.edu:8090/pub/1999–66.
6. ANSI/NISO z39.50-1995: Information retrieval: application service definition and protocol specification, 1995. URL: http://www.niso.org/standards/resources/Z39-50.pdf.
7. ZING–z39.50 international: next generation. URL: http://www.loc.gov/z3950/agency/zing/.
8. Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0, Jun 2002. URL: http://www.openarchives.org/OAI/2.0/openarchives protocol.htm.
9. ANSI/NISO z39.88-2004 the openURL framework for context-sensitive services, 2004. URL: http://www.niso.org/standards/resources/Z39_88_final_ANSI pending.pdf.12
10. Faraday, M. 1826, quoted in J.G. Crowther, *British Scientists of the Nineteenth Century*, Routledge and Kegan Paul Ltd, London, 1935, p. 96..
11. Atkins, D. *et al.* National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, Revolutionizing Science and Engineering through Cyber-infrastructure, 2003. URL: http://www.communitytechnology.org/nsf_ci_report/.
12. Newman, H.B., Ellisman, M.H. and Orcutt, J.A. Data-intensive e-science frontier research. *Commun. ACM*, 46(11):68–77, 2003. URL: http://doi.acm.org/10.1145/948383.948411.
13. Berners-Lee, T. Semantic web road map, Sep 1998. URL: http://www.w3.org/DesignIssues/Semantic.html.
14. arxiv e-print archive. Started in 1991, as of Mar 2005 it has 310,000 submissions and is expected to receive 4,500 new submissions in 2005. URL: http://arXiv.org.
15. Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C. and Warner, S. Rethinking scholarly communication: building the system that scholars deserve. *D-Lib Magazine*, 2004, 10. URL: http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html.
16. Brody, T. Citebase search and citation analysis. URL: http://citebase.eprints.org/.

Simeon Warner
Cornell Information Science
Ithaca, NY 14850, USA
Email: simeon@cs.cornell.edu
Website: www.cs.cornell.edu/people/simeon