Available online at http://www.ges2007.de

A peer-to-peer network to support scholarly communication

L. Seidenfaden, B. Ortelbach, S. Hagenhoff and M. Schumann

Institute for Information Systems, Universität Göttingen,
Platz der Göttinger Sieben 5, 3707, Göttingen, Germany
Email: {lseiden|bortelb2|shagenh|mschuma1}@uni-goettingen.de
phone: (+49 0551) 39-4442, fax: (+49 0551) 39 9735

Abstract

The number of scientific journals and thereby the number of published articles grew with an enormous rate in the last century (e.g. Price 1986; Henderson 2002). In the second half of the 20th century the system seemed to abut against its boundaries, because in relation to research budgets, library budgets did not grow fast enough to cover all the scientific output produced. Price increases well above the inflation rate set by commercial publishers that bundle disproportionately high market power – especially for journals in the Science-Technical-Medicine-Sector in the last thirty years – intensified the situation even further. This situation is known as the serial crisis. New Information and Communication Technology (ICT) driven publication models are established and seem to be a promising way out of the crisis because they reduce distribution costs significantly. Especially the open access (OA) movement that advocates free electronic access to scientific output is subject to a fierce public debate. In this paper we will detail problems associated with OA and suggest a Peer-to-Peer (P2P) system that supports electronic scholarly communication as a tool to address the economic problems mentioned above.

1 Introduction

The number of scientific journals and thereby the amount of published scientific content increased enormously in the last century (e.g. Price 1986; Henderson 2002). Proportionately to the research budgets, the budgets for libraries did not grow fast enough to cover all the scientific output that was produced. As a result, the area-wide adequate supply with scientific literature could not be sustained. Price increases well above the inflation rate (Bergstrom 2001; Orsdel/Born 2003) set by commercial publishers in an almost monopolistic market – especially for journals in the Science-Technical-Medicine-Sector in the last thirty years – intensified the situation even further. In the literature this situation is called *serial crisis* (Woodward/Pilling 1993). It is regarded as one of the driving forces that lead to changes in the system of scholarly communication.

In this context information and communication technologies (ICT) offer new possibilities to maximize the access to research results (e.g. Harnard/Brody 2004). As the technological enabler these technologies are the foundation for electronic publishing and business models which are necessary to handle the dissemination of scientific information more efficiently than the current system does and are likely to have impact on the traditional value chain of scholarly communication, e.g. the Open Access (OA) movement that advocates free electronic access to scientific literature in the established system of scholarly communication. There are two ways of achiving OA (BOAI 2006; Guédon 2004; Harnad et al. 2004; Bolman 2003):

(1) The “gold road”, in which the authors publish their work by (for or not-for-profit) open access publishers that charge an author fee to cover publication costs but make the content freely available to users.

(2) The “green road”, in which the authors themselves archive an electronic copy of an article previously published in a traditional (i.e. subscription based) journal, in a repository (Beier/Velden 2004; Crow 2002) or on their own homepage freely accessible to the public.

Although a variety of OA models that combine different publication and revenue schemes have been developed, this paper focuses on the two idealisitc types mentioned before as most new models can be linked to them in some way.

2 Economic weaknesses of OA

As mentioned before, the OA Model seems to be a promising solution for the serial crisis. However, the success of the OA movement can be questioned for a couple of reasons. These issues can be identified on different levels:

(I1) From a media economics perspective, the market power is simply shifted to other players (i.e. open access publishers), leading to increasing author or membership fees in the “gold road” model instead of subscription prices in the traditional mode (Frank et al. 2004). Therefore it is questionable that OA really is cheaper than the traditional model since payment streams are simply redirected but the costs still occur (Bolman 2003; McCabe/Snyder 2004).

(I2) From a media management perspective, the business models of open access publication forms are of interest. The different variations of author-pays-models are not tested towards their sustainability and several voices question that the fees charged so far are sufficient to cover publication costs. The not-for-profit OA Publisher Public Library of Science for example recently increased its author fees by 66%. In addition, the break even point and therewith the financial success of an author-pays journal heavily relies on the rejection rate i.e. scientific quality, because processing cost for the publisher increase linear with the number of articles reviewed but rejected and therefore not published. Thereby, lowering the rejection rate allows charging lower author fees which means that an economic factor is intertwined with the scientific aim of the journal that may lead to lower scientific quality (Bolman 2003; McCabe/Snyder 2005). Furthermore, an author-pays model may put financial burdens on research institutions with two possible results: (1) only scientists belonging to wealthy institutions can publish and (2) institutions that generate high research output have to face disproportionate financial burdens. In addition, authors are generally not willing to pay high publication fees, which leads to a lack of acceptance of open access (Cozzarelli et al. 2004).

(I3) From an information systems perspective, the lack of standardisation in the self-archiving model (“green road”) when authors put their contributions on their own website is a problem for implementing open access because it reduces the awareness for newly published works by restricted searchability. Furthermore, archiving and retrieval of electronic copies cannot be ensured. This can be circumvented by putting a copy in interoperable (i.e. OAI-PMH compatible) electronic repositories. But so far, authors rather self-archive on their own website than in (institutional) repositories (e.g. Swan/Brown 2005) which makes it necessary to find a means that allows standardized self-archiving without having scientist forced to place copies in electronic archives by mandating-policies.

3 Functions of scholarly communication and user requirements towards scholarly communication means as the determinants of the P2P systems' functionality

The application of a P2P system in the scholarly communication is more complex than in e.g. file sharing networks for recordings as in scholarly communication it needs to fulfil the four generic functions defined by Kircz/Roosendaal (1996) like any other scholarly publication mean:

(1) the *registration* function that relates research results to a particular scientist who claims priority for them,

(2) the *certification* that concerns the validation of research,

(3) the *awareness* function that leads to disclosure and search needs and

(4) the *archiving* function that concerns the storage and accessibility of research results.

It is obvious that not all functions are provided as necessary in current P2P filesharing applications, e.g. there is no formal quality assurance mechanism in those systems that insure the quality of the content. The implementation of quality assurance is important as the scientists requirements towards communication means show. The requirements towards the P2P system are derived from various studies that examine publication and reception behaviour of scientists (Schauder 1994; Swan/Brown 1999; Swan/Brown 2003; Rowlands 2004). The main requirements are:

(R1) Easy Access: The system need to provide convenient access to its contents.

(R2) Standardization: The system must support (metadata) standards (e.g. OAI-PMH) that allows search engines or A&I services to discover the publication and therewith increase the awareness, i.e. the visibility of articles published.

(R3) Topic specific communities: The need to define communities of scientists with similar research interests in the system can be derived from various studies.

(R4) Quality assurance/reputation: Furthermore, the necessity to establish efficient certification mechanisms that ensure the scientific quality of the works published is a consequence of the fact that unlike other industries, in scholarly communication the reputation of the author is directly bound to the distribution channel.

(R5) Long term archiving: It proved as important to authors, that their work is archived for a long time and thereby accessible to future generations.

The functions of scholarly communication and the requirements are used to define the P2P systems functionality. The following description of the system focuses on the functionalities that support scholarly communication.

4 P2P in scholarly communication

4.1 P2P

4.2 Basics

In this subsection P2P network architectures are described and categorized. Miller (2001) characterized P2P networks by five key properties:

- The network facilitates real-time transmissions of data between the peers.
- Peers can function as a client and a server.
- The primary content of the network is provided by peers.
- The network gives control and autonomy to the peers.
- The network accommodates peers that are not continuously connected.

P2P networks are not structured the same way, in fact a lot of degrees of freedom exist while constructing such a network. A classification of existing systems (Hong 2001) should be followed that differentiates three classes (see figure 2):

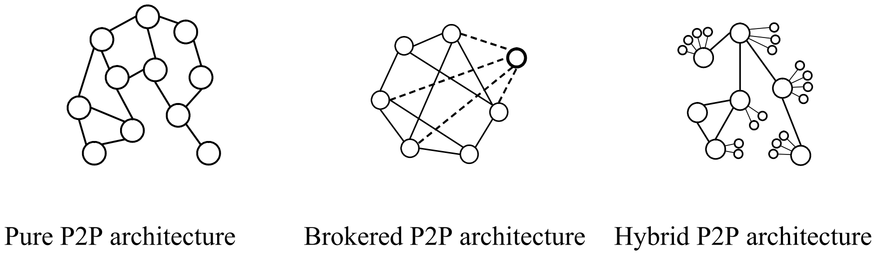

Figure 1: P2P-Architectures

In a pure P2P structure there is no central unit for coordination purposes which leads to unreliable search behaviour and performance issues. To circumnavigate the issues regarding performance and scalability, the brokered architecture is coordinated by a central server. This ensures a faster discovery of peers and content. However, the server does not provide resources such as content or disk space; it only provides coordination mechanisms. Pure or brokered architectures do not mark alternative concepts. It is possible and often reasonable to combine both within a hybrid architecture in order to bring the advantages of complete decentralisation and a central unit together. Independent of its architecture, a P2P Network can be organized in a structured or unstructured manner. Unstructured networks e.g. Gnutella, while not centrally planned in structure, grow according to a simple self-organizing process (Adamic et al. 2002). In contrast, in structured network protocols e.g. chord maintain a certain logical structure ("overlay") regardless of the size and the type of the (underlying) network (Stoica et al. 2001, Dabek et al. 2001) which improves the information retrieval.

4.3 How P2P may address the weaknesses of OA

In order to address the issues raised above, we developed a highly distributed P2P application system for the dissemination of research results that is build upon the idea of P2P filesharing applications which are the most efficient method for the distribution of music and recordings ever used. P2P might be an interesting approach because scientists strive to gain the widest dissemination of their research findings to add up reputation which is the prerequisite to get promotion. Albeit the efficiency of distribution it has to be stressed that filesharing in the music industry is commonly illegal. However, we focus on a legal application for P2P. This is possible in the wake of Open Access because the authors have the right to "give-away their work" as Harnard puts it (see http://www.ecs.soton.ac.uk/~harnad/Tp/ariadne.htm) using e.g. the creative commons licence. This is also true if the P2P system would be used for self archiving since close to 90% of the publishers allow self-archiving by the authors. Thereby the usage of the P2P application would be legal.

The reasons to use the P2P paradigm to address the economic issues raised above are explained before the functionality of the system is described in detail. From an economic perspective, the system addresses the weaknesses of OA in the following ways:

(I1) A possible solution to the problem of market power concentration seems to be disintermediation i.e. distributing the functions of intermediaries over all parties involved rather than on one player. Distributing the functions of scientific communication to the users of the system will eliminate publishers or other intermediaries from the value chain. The client application that runs on the scientists' computer, will share its resources (e.g. papers) with other members of the scientific community, no publication fees will be charged for publication. This would not change very much as scientists already fulfil the roles of authors and reviewers without financial remuneration.

(I2) Thereby, the question who pays for the intermediary (being it subscriptions or author fees) and therefore the question for a sustainable business model for it becomes redundant.

(I3) On technical level, the problems regarding the heterogeneous forms of self archiving on the "green road" can be addressed by developing an application that is compatible to various standards (e.g. OAI-PMH, Dublin Core) and a stable archiving structure. This would lead to the development of complementary products (personalized search tools, linking services) that increase the usefulness of the P2P network.

4.4 Functionality of the P2P system

With regards to the fundamental user requirement R1, it seems to be advantageous to combine the publishing and searching in one client software i.e. providing single point of access to the system rather than having different tools for publishing and searching electronic scholarly content.

The system is organized in topic specific communities that bundle scientists with similar research interests according to R3. Within these communities three user levels are distinguished in order to organize the community and the quality assurance (R4): users are only entitled to use content and do no need to register. Authors need to be registered to ensure there identity and can submit papers. Referees are permitted to perform the review of papers and accept/reject them they also need to register.

In the following the functionality of the suggested P2P system is explained using a simple layer model that clusters task specific functionality and reduces complexity (see figure 2). The layers are derived from the four functions of scholarly communication.

Registration is fulfilled in the registration layer where a timestamp is added to every submitted document and it needs to be signed using the authors' private key. Furthermore, if necessary, the review process is to be started (R2).

The *archiving* layer handles the decentralized archiving of submitted documents (R5). Therefore, an efficient distribution and retrieval mechanism for the documents stored is to be implemented here. We will not deal with issues of long-run archiving since this is not the focus of the paper (see e.g. Reich/Rosenthal 2001), but use the approach described in (Gehrke/Seidenfaden/Baule 2005) as it shows an efficient way of archiving information in a P2P system.

The *awareness* layer provides mechanisms for the notification of the relevant scientific community, i.e. a list of the most recently submitted papers is displayed in the client software and users that have subscribed to the email notification service are notified via email. Furthermore, besides rudimentary search mechanisms of the system the layer is able to allow searching on peers from external search engines (e.g. Google Scholar; Scirus) by providing OAI-compatible interfaces.

The *certification* layer supports the review process, whose functionality in contrast to the functionalities aforementioned cannot be realized solely by technical means because it involves human knowledge and judgement (R4). After a paper is submitted by an author, the layer forwards it to randomly chosen individuals that have previously registered as referees in the specific community. One individual of the referees is chosen to organize the review for that paper and to notify the author of its results.

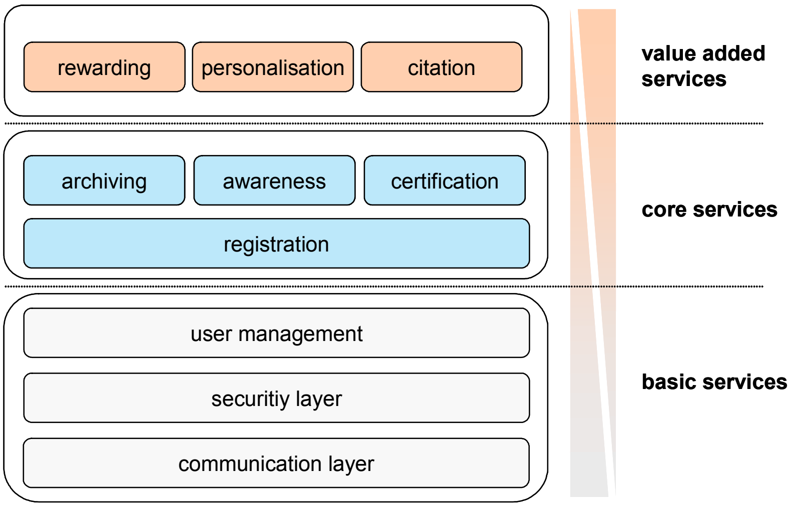

Figure 2: Layer model of the P2P-prototype

By fulfilling the functions and user requirements the system allows easy publishing and self archiving for authors in a standardised application system (*registration*) and ensures the *awareness* by obeying common meta-data standards for the system wide search functionality. Furthermore, external search engines e.g. OAIster can link to the system. By saving redundant copies of works on a large number of clients, the *archiving* can be fulfilled by the P2P application (on a very basic level). In addition libraries can connect to the system and provide archiving services. *Certification* mechanisms are integrated in the decentralised system in order to ensure the quality of the content. This is done by establishing groups of peers that share the same interest (or research topic and therefore have the knowledge necessary for review). Thereby, no economic factor such as author fees will play a role during the review process. Furthermore, functionality for annotations (e.g. for open peer review) is provided.

5 Dissemination of the P2P system and usage scenarios

For the introduction of the system in academia it may be sensible to consider a two stage approach.

At first, the P2P-system can be used as an efficient self-archiving tool for content primaryly published in traditional subscription based journals. This use case supports the green road to OA. The dissemination of the p2p system can occur via scientific societies that recommend the system to its members and thereby creates a substantial user base. With regards to the fundamental user requirement "easy access", the most important factor for using the network is that the author can publish its contribution once and it can be automatically included e.g. in a libraries archiving service and/or the authors homepage (because the P2P-System is OAI-PMH compliant). This can lead to greater acceptance of Open Access on the author's side (who are presently not willing to self archive; only 15-20% of the papers are put into repositories). Furthermore, the impact is increased i.e. more citations because the paper is openly available (see e.g. Antelman 2004; Harnard/Brody 2004). This helps to build up trust in the system which is necessary to be accepted as a publication form in the scientific community.

Second, when the system is trusted, it may be possible to eliminate costly publishers from the value chain and use the system as the primary publication tool to ensure the efficient distribution of research papers and data. Although the primary research would be free to use, it is possible that publishers link to the system and provide value-added services (personalization, enhanced certification, overlay journals) for which users have to pay (see figure 3).

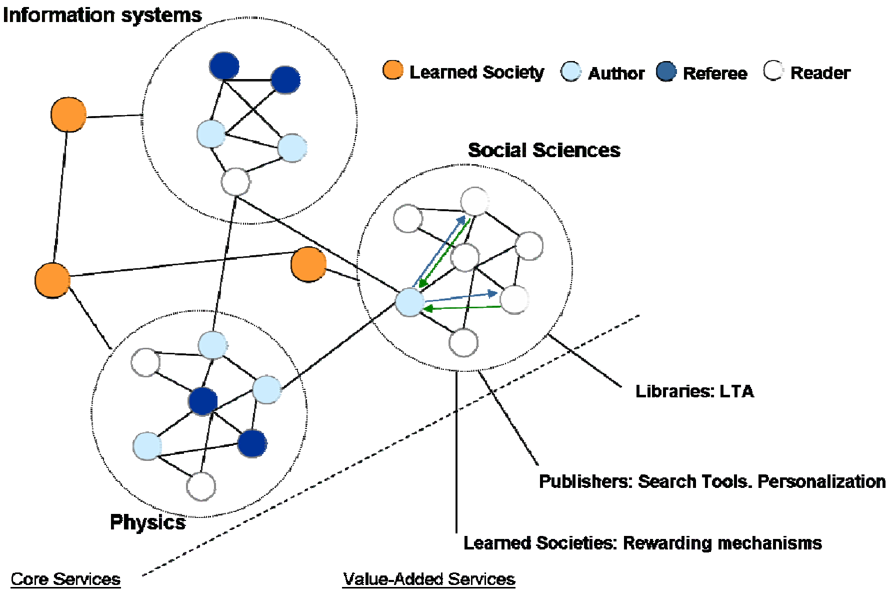

Figure 3: Interaction of the P2P-System and 3rd parties

6 Conclusion

The paper developed a concept and a prototype for the distribution of electronic scholarly articles over a highly decentralized network. It is shown that a P2P network is generally able to support electronic scholarly communication. In contrast to the existing system of scholarly communication a P2P network distributes market power among its participants rather than bundling it on one player. Thereby, it exploits the fact that today scientists already work without financial remuneration in the scholarly communication system (as authors and referees) and bundles their roles in the client software. As a result, only one client application is necessary to fulfil both roles, in contrast to the existing system which requires different tools for publishing and information retrieval. However, the system of scholarly communication is not going to be changed fundamentally (i.e. making publishers redundant in short or middle term) by this system, but it provides a useful tool that has the potential to advocate open access on the green road by providing easy and standardized self archiving by authors and easy access and convenient search ability by OAI-PMH-compatible search means and thereby changing the structure of scholarly communication in the long run. We completed a prototypical implementation of the system that demonstrates that a highly decentralized approach is technically feasible. The prototype is now in operation at our institute and scientists share there working papers through the system. Thereby the systems' functionality is tested in a first approach.

References

Antelman 2004: Antelman, K.: Do Open-Access Articles Have a Greater Research Impact? in: College research libraries, 2004, 65, 5, pp. 372-383.

Beier/Velden 2004: Beier, G./Velden, T.: The eDoc-Server Project: building an institutional repository for the Max Planck Society. High Energy Physics Libraries Webzine, 9, URL: http://library.cern.ch/HEPLW/9/papers/4/.

Bergstrom 2001: Bergstrom, T. C.: Articles - Free Labor for Costly Journals?. The journal of economic perspectives, 15, 4, pp. 183-198.

BOAI 2006: The Budapest Open Access Initiative, http://www.soros.org/openaccess/.

Bolman 2003: Bolman, P.: Open Access: marginal or core phenomenon? Information Science and Use, 23, pp. 93-98.

Cozzarelli 2004: Cozzarelli, N. R.: EDITORIAL - An open access option for PNAS. National Academy of Sciences Washington, DC. In: Proceedings of the National Academy of Sciences of the United States of America, 101, 23, p.8509.

Crow 2002: Crow, R.: The case for institutional repositories: a SPARC position paper, Washington, DC.

Frank et al. 2004: Frank, M., Reich, M., Ra'nan, A.: A Not-for-Profit Publisher's Perspective on Open Access. Serials review, 30, 4, pp.281-287.

Gehrke/Seidenfaden/Baule 2005: Gehrke, N., Seidenfaden, L., Baule, R.: Statistical Basics of a Reliable World-Wide Peer-to-Peer Memory, Proceedings of the Eleventh Americas Conference on Information Systems, Omaha, Nebraska, USA, 3.-6. August 2005.

Guédon 2004: Guédon, J.: The Green and Gold Roads to Open Access: The Case for Mixing and Matching, in: Serials review, 30 , 4, pp. 315-328.

Harnad et al. 2004: Harnad, S./Brody,. T./Vallières, F./Carr, L./Hitchcock,. S. G. Y./Oppenheim, C./ Stamerjohanns, H./Hilf, E. R.: The Access-Impact Problem and the Green and Gold Roads to Open Access. In: Serials review, 30 , 4, pp. 310-314.

Harnard/Brody 2004: Harnard, S., Brody, T.: Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, in: D-Lib Magazine, 6/2004, 2004, 10, 6.

Henderson 2002: Henderson, A.: The Growth of Printed Literature in the Twentieth Century. In Abel, R. E., Newlin, L. W. (Eds.): Scholarly publishing: books, journals, publishers, and libraries in the twentieth century, New York, pp.1-23.

Kircz/Roosendaal 1996: Kircz, J., Roosendaal, H. E.: Understanding and Shaping Scientific Information Transfer, Joint ICSU Press/UNESCO Expert Conference on ELECTRONIC PUBLISHING IN SCIENCE, Paris, 19.-23.2.1996.

McCabe/Snyder 2004: McCabe, M. J., Snyder, C. M.: The Economics of Open-Access Journals, Mimeo, Georgia Institute of Technology.

McCabe/Snyder 2005: McCabe, M. J., Snyder, C. M.: Open Access and Academic Journal Quality. American Economic Review, 95, 2.

Price 1986: Price, Derek J. de Solla: Little Science, Big Science and Beyond, Columbia University Press, New York.

Reich/Rosenthal 2001: Reich, V., Rosenthal, D. S. H.: LOCKSS: A Permanent Web Publishing and Access System, DLIB, 7.

Rowlands et al. 2004: Rowlands, I., Nicholas, D., Huntington, P.: Scholarly Communication in the Digital Environment: What Do Authors Want? Learned Publishing, 17, 4, pp.261-273.

Schauder 1994: Schauder, D.: Electronic Publishing of Professional Articles: Attitudes of Academics and Implications for the Scholarly Communication Industry. American Society for Information Science: Journal of the American Society for Information Science, 45, 2, pp.73-100.

Swan/Brown 1999: Swan, A., Brown, S.: What authors want: the ALPSP research study on the motivations and concerns of contributors to learned journals.

Swan/Brown 2003: Swan, A., Brown, S.: Authors and electronic publishing: what authors want from the new technology. Learned Publishing, 2003, 16, 1, pp. 28-33.

Woodward/Pilling 1993: Woodward, H./Pilling, S.: The International Serials Industry: an overview. In: Woodward, H./Pilling, S. (Eds.): The International Serials Industry, Hampshire, pp. 1-22.