Dangzhi Zhao[†] **and Andreas Strotmann**[‡]
[†]University of Alberta, Edmonton, Alberta, Canada
[‡]University of Cologne, Cologne, Germany

Towards a Problem Solving Environment for Scholarly Communication Research

Abstract: We explore the need for and the architecture of a Problem Solving Environment (PSE) for scholarly communication research through an examination of the problem area it addresses, the problem solving processes that scholarly communication researchers commonly employ, and the existing data sources for this type of research.

As scholarly activities in many fields become increasingly heterogeneous and complex, and as information technologies become more and more powerful, research on how to better facilitate scholarly activities with these powerful technologies is also increasing. This explains why the design and implementation of Problem Solving Environments (PSEs) has become a research focus in computational science.

It has been predicted that PSEs will eventually become "the main gateway" for scholars to access resources and will "change the pervading research culture, making it more open and accountable" through their support for "transparent access" to heterogeneous distributed resources and through the "seamless integration" of new software, hardware and data resources (Walker and Rana 2003, 6).

Here, we explore a PSE for scholarly communication research, aiming to contribute to the study of scholarly communication, to help improve scholarly communication systems, and to advance research on PSEs. We first discuss the need for a scholarly communication research PSE, and then propose an architecture that follows naturally from the requirements derived by examining the problem area it addresses and the problem solving processes that scholarly communication researchers commonly employ.

The present study is based on our experience with a large scale citation analysis study (referred to in this paper as "the example study") that compares the scholarly communication patterns revealed from research papers published on the Web as indexed by *ResearchIndex* with those demonstrated in print journals as indexed by *Science Citation Index (SCI)* in the XML research field (Zhao 2003).

1. What is a Problem Solving Environment?

Like many terms in Information Science, different people use the term PSE in different ways (Abrams et al. 2003). Although the term is hard to pin down exactly, there have been attempts at a definition. The following one, by Gallopoulos, Houstis, and Rice (1994, 11-12), appears to be the most frequently quoted in research papers on PSEs:

> A PSE is a computer system that provides all the computational facilities necessary to solve a target class of problems. These features include advanced solution methods, automatic or

> semiautomatic selection of solution methods, and ways to easily incorporate novel solution methods. Moreover, PSEs use the language of the target class of problems, so users can run them without specialized knowledge of the underlying computer hardware or software. By exploiting modern technologies such as interactive color graphics, powerful processors, and networks of specialized services, PSEs can track extended problem-solving tasks and allow users to review them easily.

Thus, a PSE applies state of the art technology to create powerful and convenient computing environments for specific application areas (Abrams et al. 2003). A PSE provides software tools and expert assistance to scholars in a user-friendly environment. With a PSE, a scientist can concentrate on the science rather than spending too much time and energy on "inessential and often arcane details of specific hardware and software systems" (Cheng 1996), resulting in higher research productivity.

Although research on PSEs began in the 1960's, lack of computing power limited both the research on, and practice of, PSEs in the 1970's and 1980's (Cheng 1996). Since the 1990's, however, research and practice of PSEs have been very active due to advances in key technologies and due to stronger motivations.

The ability to build good PSEs has increased with the power and sophistication of the key technologies such as networking, visualization, parallel processing, and interface tools. The need for PSEs has increased with the complexity and heterogeneity of the data, the applications and the problem solving processes that researchers employ. Without a sophisticated and usable PSE, it is often virtually impossible to solve, e.g., large-scale engineering design problems that are extremely heterogeneous, such as multidisciplinary design optimization used in aerospace engineering (Abrams et al. 2003).

2. Why a Problem Solving Environment for scholarly communication research?

The essence of a PSE is thus to free researchers from inessential technical problems that can be extremely time consuming, by making available a full set of high level tools and resources in a user-friendly environment. This allows researchers to concentrate on their research problems, and, more importantly, supports more sophisticated studies and research collaborations that may be very difficult if not impossible without a PSE. This is especially true in scholarly communication research, which can be clearly seen from studies of scholarly communication using citation analysis approach, one of the most powerful and best-known approaches to the study of scholarly communication.

2.1. Limitations of existing data and tools for citation analysis studies

Citation analysis has been successfully applied in the study of formal aspects of scholarly communication, such as the study of the characteristics and the evolution of scholarly communities and networks, the evaluation of scholarly contributions, and the study of the diffusion of ideas (Borgman 1990). It has contributed to the study of informal scholarly communication by helping sociometric studies to identify scholarly communities or specialties, to discover interesting structures and processes in scholarly communication for deeper study, and to validate the results of such studies (Zhao 2003).

However, citation analysis studies have largely been limited to what can be done with the databases produced by the Institute for Scientific Information (ISI) that have for a long time been the only viable data source for citation analysis studies.

The ISI databases have significantly contributed to the advancement of citation analysis theory and methodology, but have also drawn considerable criticism and limited the development of citation analysis theory and methods (Smith 1981; MacRoberts and MacRoberts 1989). Among the major problems caused by the dependence on ISI databases are the ways in which these collections treat multiple authorship, and their limited and biased coverage.

The ISI databases index only the first authors of cited documents, which adversely affects the accuracy of citation and co-citation counts in citation analysis studies and, as a result, limits the applicability of citation analysis in the evaluation of scholarly contribution and in the study of intellectual structures of disciplines. Although in principle there are ways to overcome this problem, in practice it is not easy to do so especially for large-scale studies, but the expansion of the ISI databases over time alleviates the problem as more and more cited items fall within the databases as source papers (citing papers).

The ISI databases only cover journal articles selected by the ISI's advisory boards of experts in each topic represented (ISI 2000). The highly selective choice of journals may have caused validity problems in citation analysis studies. The "journal only" coverage does not admit citation analysis studies of other types of publications such as conference papers and technical reports. As a result, studies are often limited in scale and therefore of limited validity. This is an increasingly serious problem as new publishing formats emerge through the rapid development of information technology. Biases in the coverage of the ISI databases have also been identified in terms of the language of papers. Most papers indexed there are in English, which means that the ISI databases cannot support the study of scholarly communication patterns demonstrated in other languages, nor the comparison of research communities in different countries.

The number of articles in top Information Science journals discussing techniques for retrieving citation and co-citation frequencies from the ISI databases is a clear indication that citation analysis studies using the ISI databases can be extremely difficult and time consuming due to their coverage and structure, such as the severely limited information available on cited items. It has therefore been a dream of citation analysts to have a tool at their disposal that would integrate the many steps involved in citation analysis into "one smooth-flowing, economical machine process" (White 1990, 104).

With the Web becoming a powerful communication medium, full text research papers with complete reference lists are increasingly becoming available on the Web. Search engines and even citation indexes such as ResearchIndex are emerging to help researchers make full use of these resources. Compared to the ISI databases, data sources and tools on the Web cover a wider variety of document types such as conference papers, technical reports, degree theses, and preprints in addition to journal articles, and provide more complete information about cited papers such as titles, all authors, full source names, and citation context. These may facilitate a larger variety of inquiries and more sophisticated methods that may be very difficult, if not impossible, with the ISI databases, such as comparative analysis between document types, citation context analysis and weighted citation counting (Clever 1999; Narin 1976). However, these data sources and tools currently have their own problems (Zhao 2003).

One problem is that although more and more research papers are being made available on the Web, citation indexes for these publications are still very limited in

number and in disciplines covered. One such citation index has been developed by the Open Citation Project for the LANL e-print archive that is maintained at the Los Alamos National Laboratory as a digital library for physics, mathematics, computer science, and related disciplines. Another index, ResearchIndex, is currently limited to computer science although its technology could be adapted to other fields (Lawrence et al. 1999).

Another problem is the data format. These Web citation indexes were not developed with the goal of exchanging data with other services or applications in mind. As a result, they provide no option for retrieving search results in a standardized format that is independent of their presentation, so that citation analysis researchers or other Web services could rely on such an unchanging format to analyze the data retrieved. Search results from ResearchIndex are only available in HTML format, for example, for cited papers that are not in its database. Consequently, collecting data from these databases for scholarly communication research is very difficult, because their presentation format constantly changes, and cues available in HTML for distinguishing different data elements such as author, title and publishing venue are very limited. Moreover, ResearchIndex indicates the total number of documents that meet search criteria but provides at most 1,000 of these documents to the user. The limit placed on the number of retrieved papers certainly limits the scale of ResearchIndex based citation analysis studies, and therefore the usefulness of ResearchIndex for citation analysis studies.

A third problem is associated with the automated extraction of citations from publications on the Web by citation indexes. Since papers published on the Web can be in all kinds of formats using various referencing conventions, fully automatic citation indexing tools like ResearchIndex tend to mix up information about authors, titles, sources, etc. when uncommon or non-standard writing and referencing formats are encountered. It is also very difficult for them to correctly capture publication date information, which can be either absent or hidden anywhere in the document. As a result, these tools do not support time-related studies, such as the evolution of scholarly communities or the diffusion of ideas over time, and may have validity problems just as the incompleteness and biases of the ISI databases have been discussed in the literature as potential validity problems.

2.2. The need in scholarly communication research for a wide range of data sources

Scholarly communicative activities are increasingly being conducted over the Internet, in electronic journals and books, online conference proceedings, discussion boards, newsgroups, chat rooms, Web-based group support systems, just to name a few. This provides new material and data for scholarly communication research that may facilitate a larger variety of inquiries such as how scholarly communication is being transformed, what the similarities or differences between the new formats and the traditional ones and between different document or language types may be, and how the new formats facilitate or inhibit scholarly communication processes.

It seems natural for scholarly communication researchers to study such materials as a data source for citation analysis studies, but, we have only seen a few studies so far that make use of Web data sources. One reason for this is clearly the difficulty involved in using these data sources as discussed above. Other reasons may include some of the concerns citation analysts have about the use of these data sources, including: (1) Web-publishing is not as well controlled as journal publishing and therefore might be viewed as being flawed for citation analysis; and (2) Citation analysis of ISI data is believed

enough to see scholarly communication patterns in a research field as papers indexed in these databases are thought to be "the most important" part of the literature in the field.

Zhao (2003) addresses these concerns through a citation analysis study on XML research using both data on the Web as indexed by ResearchIndex and a traditional citation analysis data source (i.e. ISI databases). We found that in the XML research field author rankings by number of citations resulting from the two data sources were highly correlated when the same citation counting method was used, while different intellectual aspects of that research field were revealed from the two data sources. This indicates that the evaluation of scholars based on the collective view of authors as demonstrated in the material on the Web should be considered as equally valid as that in the ISI databases, provided the discipline being studied is well-published on the Web. In order to gain a complete picture of the scholarly communication patterns, multiple data sources should be used rather than only the ISI databases or Web data sources such as ResearchIndex.

2.3. Scholarly communication research improving scholarly communication systems

Citation analysis not only helps understand scholarly communication structures and processes but also aids information retrieval (which is, in effect, one type of communication) (Borgman and Furner 2002). The ISI databases and *ResearchIndex* have demonstrated the value of incorporating citation analysis results into information retrieval systems by providing such information as the number of citations each document receives, co-cited documents, and various indicators of journal quality. There are other possibilities of making use of citation analysis in information retrieval – evaluative citation analysis results can help retrieve high quality documents and publications by core players (authors, institutes, etc.), and relational citation analysis results may help expand queries and refine searches through clustering of documents, authors and sub-areas. Moreover, citation networks presented at various levels (micro or macro) by applying advanced scientific visualization techniques can assist literature review, and information management (Lin, White, and Buzydlowski 2001).

Although it is not difficult to understand the benefits of citation analysis results in IR systems, the application of citation analysis in this area has not been fully explored, partly because most IR systems deal with data that does not include references. As full-text documents that do include references, and autonomous citation indexing tools that use them, are becoming available on the Web, the obstacles imposed by data sources are disappearing, and such applications are not only feasible but also finding more potential within the Web infrastructure. For example, integrating citation analysis facilities into other Web services such as digital libraries and search engines could help users determine the relevance, quality and interrelatedness of scientific papers they encounter while using these services. Some efforts already exist that make use of citation links and citation analysis within the Web infrastructure to provide value-added information services such as relevance ranking in search engines (e.g. Google), reference linking (e.g. CrossRef and OpenURL) and citation indexing (e.g. ResearchIndex and the Open Citation Project).

In summary, scholarly communication research needs to access a wide range of data sources, which is not supported by existing hard-to-use citation indexes. As a result, scholarly communication researchers have been limited to very few data sources, spending a huge amount of their research time dealing with the difficult tools. The potentials of applying scholarly communication research to improving scholarly communication systems have not been fully realized. A PSE for scholarly communication

research is thus necessary that, by applying current technologies, provides transparent access to heterogeneous data sources and tools to facilitate studies of a wider range of problems, integrates the separate steps of scholarly communication research into "one smooth-flowing, economical machine process" (White 1990, 104), and helps applying scholarly communication research to networked information organization and retrieval.

3. What should be included in a PSE for scholarly communication research?

A PSE for scholarly communication research would be a computer system that uses current technologies to provide all the computational facilities necessary to carry out scholarly communication research. We will therefore first define the scope of scholarly communication research and then identify some of the computational facilities necessary for conducting scholarly communication research.

3.1. Scholarly communication research

Scholarly communication research is "the study of how scholars in any field (e.g. physical, biological, social, and behavioral sciences, humanities, technology) use and disseminate information through formal and informal channels" (Borgman 1990, 13). Communication plays an important role in scholarship. It is through communication that scientific discoveries become known and, as a result, impact the development of science. It is also through communication that scholars' contributions get evaluated and recognized by peers. In other words, communication is the basis of the reward system of science which provides incentive for research (Merton 1979).

The importance of communication in scholarship has drawn attention from a diverse group of researchers including sociologists of science, communication researchers, historians of science and information scientists, who have contributed to the study of scholarly communication from different perspectives for different purposes. Some have attempted to build models of information flow in scholarly communication in order to improve the effectiveness and efficiency of information systems (Garvey 1979), whereas others have tried to understand how social organization among scientists inhibits or facilitates scholarly communication processes (Crane 1972). Some have examined the structure and processes of scholarly communication through formal channels such as the journal (Small and Griffith 1974), while others have explored what is accomplished by the circulation of scientific information on an informal basis (Crawford 1971).

While the study of scholarly communication is broad in scope, communication behavior can be divided into formal and informal domains (Crawford 1971). Formal scholarly communication is communication behavior of scholars demonstrated through scholarly publications. Informal scholarly communication takes place within an interpersonal relation, in which scholars communicate with each other on their research.

Informal communication is primarily studied using social network analysis methods based on sociometric data which record the frequency or strength of direct and indirect social choices or contacts and are often obtained through surveys or interviews, while formal scholarly communication is primarily studied through bibliometric approaches which seek "to shed light on the processes of written communication and on the nature and course of development of a discipline […], by means of counting and analyzing the various facets of written communication" (Pritchard 1969, 348). Sociometric studies are

concerned with the changing structure of communications and social relationships associated with intellectual progress in specific scientific specialties or communities whereas bibliometric studies focus on the literature and implied intellectual structures of scientific specialties. Citation analysis is the best-known bibliometric technique.

3.2. Computational facilities needed for scholarly communication research

These can be identified through an examination of the problem solving process commonly applied in scholarly communication research.

Scholarly communication research is broad in scope. Researchers have employed many techniques and methods in the study of problems in scholarly communication, e.g. social network analysis, content analysis, or citation analysis. Different methods are employed in different problem solving processes. Although a comprehensive analysis of such processes would be necessary for designing a full-blown PSE for scholarly communication research, our analysis focuses on examining in detail a problem solving process that uses citation analysis, and only briefly refers to other relevant types of studies. We still contribute to the development of a PSE for scholarly communication research because citation analysis studies account for an important part of scholarly communication research, and some other techniques and methods, such as link structure analysis of websites and social network analysis of sociometric data, are closely related to citation analysis, and other techniques in scholarly communication research are similar in structure. Moreover, a PSE is by definition open to new methods, tools and data sources, so that although a PSE based purely on the process described here would be incomplete, new tools, once identified and developed based on processes of scholarly communication research using other techniques, could easily be incorporated into an existing PSE.

Like most scientific studies, a citation analysis study of scholarly communication starts with research problem identification and research design. This is one of the most important parts of the research and demands an in-depth knowledge of scholarly communication research, an exhaustive literature search and intensive communication with peers. However, most existing PSEs do not include facilities specifically designed for support of this part of the research.

The present study therefore starts the analysis of the problem solving process from the point where a problem has been identified and a research design has been achieved.

Obtaining a set of citing papers and citations they contain. Scholarly communication studies are usually concerned with communication patterns in certain scholarly communities. Scholarly communities can be defined in different ways in citation analysis studies, such as a nation, a discipline or an institution. In addition, citation windows are often specified in scholarly communication studies. Thus, a set of citing papers needs to be obtained — along with the references they contain — that represents a scholarly community within a specific citation window. Subsequent analyses are based on how the citing authors collectively view their scholarly community as demonstrated through their citing behavior. In the case of the example study, the scholarly community we studied was the XML research field, and citing papers resulting from a keyword search in *ResearchIndex* or *SCI* and their references were used to determine the authors that represent this community.

A PSE for scholarly communication research therefore needs to make available data sources that cover a variety of scholarly communities. For citation analysis purposes, data sources should record the citing behavior of authors and should support searching, e.g., by subject, author, date, author affiliation and, more importantly, by cited items (e.g. by cited authors). Other types of study would require access to full documents, and yet others might study digitally recorded multi-party conversations.

Regardless of the data sources used in scholarly communication studies, the PSE needs to support the retrieval of search results in a standardized format such as XML. Data sources that do so are here tentatively called "citation analysis data sources."

Data sources are increasingly available on the Web, but they may or may not be citation analysis data sources in this sense. Therefore, in addition to a structure that allows easy integration of existing citation analysis data sources, wrappers that convert data sources such as the ISI databases, *ResearchIndex* or the CogPrints repositories into citation analysis data sources would be an important ingredient of a PSE. It should also have a mechanism for being informed of and accumulating citation analysis data sources or data sources that have the potential of supporting citation analysis. An online registry-lookup service and a specialized crawler such as a "harvester" within the Open Archives Initiative (OAI) framework would be an example of this, and as in OAI, a registry-lookup service would allow data sources to be registered and searched. A crawler could go out on the Web to look for data sources that are useful for scholarly communication research.

With these facilities, scholarly communication researchers would be able to use whatever data sources they find available even if these are originally not in a convenient format for citation analysis. For example, a scholar who wants to do a study similar to the example study in a research field outside computer science, would only need to provide a set of keywords and specify search criteria that define the scholarly community she wants to study. The crawler component of the scholarly communication PSE would go out on the Web to locate research papers that meet the criteria, and the wrapper component would construct a database for these papers that is similar to *ResearchIndex* but enhanced with features such as the option of downloading search results in a standardized format.

The bottom line is that a data source that a scholarly communication researcher intends to use should contain information about the relationships in scholarly communities such as citing-cited relationships. Examples include as full text research papers with reference lists in the case of citation analysis or a listserv archive recording who responded to whom at what time about what in the case of social network analysis.

Cleaning citing paper information. Once citing papers and the citations they contain have been collected, a scholarly communication researcher usually needs to do some clean-up work such as deleting any duplicates or paper entries that do not contain the specific information that is needed for that particular citation analysis study. As an example, some opinion papers do not provide any references, and the technology used for indexing some papers that do may not be able to capture the references correctly. Such paper entries do not have any value in citation analysis studies and therefore need to be deleted from the analysis. Duplicates were found in the example study to be an issue that should not be neglected. The same research paper is sometimes published several times in different formats on the Web: as an early draft, a technical report, a conference paper and a journal paper. Automatic indexing tools such as *ResearchIndex* often cannot make the distinction between the different formats of the same research papers although humans

can easily do so. As a result, duplicates are often seen in data collected through an automatic tool on the Web.

Experience from the example study showed that some clean-up work has to be carried out manually while some, such as the removal of duplicates caused by different formats of the same author names, can be done automatically with the right programs. A PSE should therefore provide tools for identifying and removing duplicates or incomplete records, but it should also present citing papers and their references to the researcher in such a way that she can make corrections or deletions effectively and efficiently. In addition, a PSE should also provide facilities that help her with information verification, such as easy access to authors' homepages or to the full text of referenced papers.

Ranking objects by citations. Scholarly communication studies that apply citation analysis techniques usually need to rank objects by their number of citations as calculated in various different ways. Objects can be papers, publishing venues (e.g. journals or websites), scholars (e.g. authors), nations, institutions, etc. Objects to be ranked can be a predefined group such as all faculty members at a library school, or they can be all objects in the scholarly community being studied that appear in the set of citing papers collected, such as all scholars who have ever been cited by these papers.

To facilitate ranking objects by number of citations, a PSE for scholarly communication research should provide both a predefined set of commonly used methods such as complete counts, fractional counts, straight counts or more sophisticated methods (Narin 1976; Clever 1999), and a way for a researcher to define her own metric. It should also support easy incorporation of novel approaches that emerge as scholarly communication research advances.

In addition, there is a well-known problem in citation analysis caused by the fact that authors may publish under different names or different authors may have identical names. This problem needs to be addressable within a PSE. Tools for determining object identities are an important ingredient of a scholarly communication PSE, such as facilities for the identification of, and access to, authors' and institutions' publications and homepages. Such tools tend to be available on the Web, so that a PSE may simply wrap such services with translation to a standardized information exchange format.

Obtaining co-citation frequencies for selected objects. Objects to be co-citation analyzed can be selected from many sources, based on which the merit of the objects can be determined. Frequently used measures include citedness, nomination by domain experts, or appearance in selective reference books such as *Who's Who*. To facilitate this selection process, assistance with surveys or interviews of domain experts, and access to reference books and other related data sources would be a useful ingredient of a PSE.

Once objects have been selected, their co-citation frequencies need to be collected. Co-citation frequency is just one way of measuring how closely two objects are related from the point of view of citers. Any methods that do the same job can be used in relational citation analysis. Several methods can be found in the literature in addition to the two methods used in the example study, such as co-citation counts of two objects divided by the sum of the citation counts of these two objects (McCain 1999). It is quite possible that new approaches will be developed with the advance of scholarly communication research (Ahlgren, Jarneving, and Rousseau 2003). Therefore, support for multiple counting methods and incorporation of new approaches is needed in a PSE.

Compilation of co-citation frequency matrixes. When the co-citation frequencies are entered in a matrix, some rows and columns may contain too many zero cells, which means that the authors or papers corresponding to those rows and columns were not co-cited with many other authors or papers, and therefore may not be good representatives of the field being studied. Such objects and their corresponding rows and columns in the co-citation matrix may need to be deleted according to some ad hoc criteria (McCain 1990), such as the one used in the example study, that is, more than 5% zero value cells. A PSE should provide a variety of well-established filters of this kind, and allow the researcher to define her own filters.

A co-citation frequency matrix is obviously symmetric with respect to the main diagonal, but the definition for the diagonal cells of the matrixes is a technical detail that the researcher needs to decide. The main diagonal could contain the number of citations each object receives, which can dominate the off-diagonal cells despite having a different interpretation than the rest of values in the matrix. There are several proposals in the literature of determining these values, e.g. to treat them as missing value, or to scale them downward based on off-diagonal values (McCain, 1990). A PSE should provide a reasonable default setting with an option for a researcher to adjust them to her needs.

Conversion to correlation coefficient matrixes. Raw co-citation counts can be very misleading if compared directly, so that co-citation frequency matrixes are usually converted to a correlation coefficient matrix, often using Pearson's r. This conversion has the advantage that a normalized measure of the similarity between two objects takes into account their entire co-citation record rather than individual co-citation counts. Statistical software packages provide routines to calculate various types of correlation coefficients, so that access to such packages using a standardized and reusable data format would be an important component of a scholarly communication PSE. However, the PSE should focus on providing access to those aspects of a statistics package that are used extensively in the scholarly communication research area, and may ignore features of such packages that are less frequently used.

Thus, the compilation and conversion process can be supported nicely by computer programs collected in and provided by a PSE, and access to existing statistical software packages would be helpful when calculating correlation coefficients. However, it is the next step that requires the incorporation of these existing packages rather than writing programs from scratch.

Statistical analyses and their results. Citation analysis is concerned with gaining "a macro perspective on a scholarly communication process through the use of voluminous datasets" (Borgman 1990, 26), and thus often needs to apply statistical analyses such as correlation, factor, and cluster analysis or multi-dimensional scaling.

Existing statistical software packages such as SPSS or SAS contain many routines that are useful for citation analysis. The difficulty with using them is that they often require specific formats for input data and then produce results in formats that are not easy to use directly in citation analysis. The example study had to use Microsoft Excel files and sometimes text files to communicate with SPSS manually: inputting matrixes from Excel to SPSS and copy/pasting, say, coordinates of author points from the output file of SPSS's ALSCAL into Excel. The plots of author-points produced by SPSS did not work properly for the example study, so that we used LaTeX to produce better two-dimensional author maps points from coordinates produced by SPSS's ALSCAL.

It would therefore be helpful to have transparent access from within a PSE to important routines in statistical packages in such a way that both the problem specification and the results of their statistical analysis can easily be exchanged with other tools in the PSE in a standardized format. Among the statistical analysis result types our example study used were factor structures from factor analysis, dendrograms from cluster analysis, and spatial representations from multi-dimensional scaling.

Note that from the PSE user's point of view, it should be irrelevant which specific statistical package runs the statistical analysis she requested. By requiring standardized formats for communicating with such a package, any such package should do as long as it provides the functionality required by the researcher using the PSE. Furthermore, we found in our example study that it may be better to leave the visualization of statistical results to specialized visualization tools rather than using the visualizations produced by the statistical packages themselves. In other words, visualization tools need to be incorporated in a PSE just as statistics packages, with all the caveats we gave for them.

Interpretation and validation. Factor analysis, cluster analysis, MDS and other statistical approaches make it much easier to identify relationships among objects by converting these relationships to the relationships among factors or clusters that are much smaller in number, and by presenting the relationship between these clusters visually. Based on this, an interpretation can then be attempted.

As McCain (1990, 441) points out, "interpretation relies on discovering what the author clusters, factors, and map dimensions represent in terms of scholarly contributions, institutional or geographic ties, intellectual associations, and the like." The depth of interpretation depends on knowledge that the interpreter has about the research field being studied or the domain being analyzed. The scholarly communication researcher needs to either have the knowledge herself or to acquire it from domain experts who are willing to be interviewed. Tools that help with the identification of, and establishment of communication with, domain experts and with conducting interviews would therefore be very helpful components of a PSE.

Similarly, there are a number of ways of validating citation analysis results as reviewed in Zhao (2003). The basic idea is to compare results from citation analysis with judgments of domain experts obtained through interviews, surveys or existing writings such as textbooks, reviews, and historical documents. Surveying tools and access to such documents are therefore necessary ingredients of a PSE.

For both interpretation and validation, the construction of object profiles can be of great help. An object profile is a concise description of the characteristics of that object. In the case of author co-citation analysis, for example, an object profile could be a description of which publications collected in the co-citation analysis an author has co-authored, and of which social or other ties this author has with other authors in the study. Once author factors or clusters are identified, each factor can be described using profiles of authors in that factor. A profile can be produced manually by the researcher by, say, going through publications and author homepages, which is how the example study did it, but it could also be generated automatically from titles, abstracts or even the full text of articles using word frequency analyses, automatic topic extraction or other more sophisticated computational techniques. A PSE could provide valuable tools for constructing such object profiles. Object profiles can not only give valuable clues as to the meanings of clusters and factors, but also aid in interviews with domain experts.

Domain experts may not know in detail what other authors in their field have done, and may need access to object profiles in order to provide their judgments of the common interests that define a cluster. This is one reason that these interviews are often "at length" (McCain 1990b, 441). An expert can make better sense of relevant information than an outsider, and object profiles provide them with the information they need in a concise format.

Conducting interviews online in an easy to use and effective environment would make it easier to acquire and interview domain experts, because geographical obstacles would disappear. There are also a number of advantages of conducting interviews online compared with over the phone, the most important one being that the interviewer and interviewee can easily share written material such as object profiles online.

Clearly, therefore, in order to facilitate the interpretation and validation process, tools for constructing object profiles such as word frequency analysis tools, tools that help with identifying and communicating with domain experts, and environments that facilitate interviewing and surveying domain experts should be included in a PSE.

The problem solving process discussed above shows that, unlike most of the domains for which PSEs have been developed which are based on sophisticated mathematical models requiring sophisticated numerical solution methods, studies of scholarly communication are characterized by the need to access heterogeneous distributed data sources, reliance on domain experts, and intensive character processing such as citation searching, cleaning, and counting, or word frequency analysis.

In summary, therefore, resources that should be included in a PSE for scholarly communication research are at least (1) citation analysis data sources that cover a wide range of scholarly communities and a variety of data types such as sources of full text research papers and newsgroups or listserv archives; (2) filters, parsers and tools for automatic topic extraction that are useful in citation data searching and cleaning, in the construction of object profiles, and in automatic labeling of object clusters or factors; (3) citation and co-citation counting tools; (4) statistical analysis tools; (5) visualization tools; and (6) survey / interview design and conduction environments.

4. Characteristics of a PSE for scholarly communication research

Abrams et al. (2003) listed the characteristics "that are desirable, to one degree or another, in a scientific PSE." For a PSE for scholarly communication research, the following characteristics from their list are especially important.

Problem-oriented. The PSE should allow the scholarly communication researcher to concentrate on solving research problems, "without having to become a self-taught expert in networks or parallel computing or the World Wide Web." (Abrams et al. 2003)

Integrated. Scholarly communication research is characterized by the need to access heterogeneous distributed data sources and tools. A scholarly communication research PSE should make these resources available and manage them in an integrated way to provide the user "a predictable and consistent environment." (Abrams et al. 2003)

Powerful. Scholarly communication research is broad in scope. In order to allow various problems of interest to be studied, a PSE for this field should make available powerful hardware and software resources and a variety of types of data sources (papers, newsgroups, etc.) that cover as many scholarly communities as possible.

Open, flexible, adaptive. Data sources are increasingly made available on the Web and new approaches to the study of scholarly communication are constantly being developed. A PSE for scholarly communication research should be open to these new resources. In addition, it should allow sophisticated users to tailor or add to its functionality. Also, it should allow other applications to integrate with their services. For example, it should allow a full text repository to use it to provide the functionalities enabled by citation indexing and analysis that include but are not limited to those being experimented with in *ResearchIndex* and in the Open Citation Project.

Graphical, visual. In addition to a graphical user interface through which users can easily communicate with the PSE in the language of scholarly communication research, scholarly communication studies often require experimenting with visualizations of research results, either citation networks or associated social networks, for example.

Intelligent. Expert assistance is very useful throughout a citation analysis study — from choice of an input set of objects, "through critiques of proposed interpretations of the results, to identification of anomalies that merit further investigation" (McCain 1990). With the advances of technology and online availability of information about objects (e.g. authors), it is likely that a PSE could supply some expert "advice" on these issues, such as automatic topic extraction of object factors / clusters through analysis of words in titles or abstracts of research papers.

5. Architecture of a PSE for scholarly communication research

From the preceding requirements analysis for a problem solving environment for scholarly communication research, and from the practical experience we gained in our attempts to automate some of the steps in an author co-citation analysis of a research field using a variety of data sources and analysis tools, certain basic characteristics of an architecture of a PSE for scholarly communication research emerge naturally: First, the researcher's interface to such a PSE should seamlessly integrate into her everyday working environment; second, crucial components of the PSE need to take the form of web-based services that handle information in an open and standardized data format; and third, it should be straightforward to augment a PSE with additional data sources or tools.

Together, these points suggest that a scholarly communication PSE be designed as a collection of Web services that the researcher orchestrates via the PSE's user interface which, again, should take the form of a web-service with a regular web interface. Unlike the user interface, however, which would use browser-oriented data formats when communicating with the researcher's Web browser, the Web services that are orchestrated via the PSE "under the hood" would be required to hook into the PSE via standardized data formats that are determined by the type of information that a Web service handles (recall that the lack of this type of Web service access for the ISI and ResearchIndex citation databases posed a major problem for our example study).

Thus, different types of Web services would use different data formats, but all of these should be standardized XML applications. A citation database, for example, could produce bibliographic data in, say, Dublin Core format, encoded as RDF, augmented with a "cited-by" relation. Statistical analysis packages such as SPSS would likewise be hooked into the PSE as a Web service with an XML-based statistical data format (recall again that the lack of such a format caused considerable problems in the example study). Other tools, including visualization tools, can be made available in this form, too.

Since there are always so many standards to choose from, translation services (known as cross-walks in the information studies literature) will be required as web services. This is especially important in the context of a scholarly communication PSE, since a multitude of standardized formats are in active use across the world for the type of information that scholars in the field need to work with in their studies.

The service registries for data sources or tools that our requirements analysis pinpointed earlier as useful for a PSE could then be realized using any of the corresponding technologies for web services such as the Web Services Description Language or the Open Grid Services Architecture, and the PSE itself, which therefore can be seen as orchestrating a host of web services for its user, could be specified using Web technologies like the Web Services Choreography Description Language.

Thus, the architecture that emerges rather naturally for a PSE for scholarly communication studies corresponds closely and nicely to the architecture that is actively being developed for the World Wide Web itself. Thus, the study of such a PSE for one particular area of information science could lead to a better and deeper understanding of the as-yet unrealized potentials that the future Web holds in store for the field.

References:

Abrams, M., D. Allison, D. Kafura, C. Ribbens, M.B. Rosson, C. Shaffer, & L. Watson (n.d.). *PSE research at Virginia Tech: An overview*. Retrieved April 9, 2003, from Virginia Polytechnic Institute and State University Web site: http://research.cs.vt.edu/pse/intro.html

Ahlgren, P., B. Jarneving, and R. Rousseau. 2003. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology* 54: 550-560.

Borgman, C.L., ed. 1990. *Scholarly Communication and Bibliometrics*, Newbury Park, CA: Sage Publications, Inc.

Borgman, C.L. and J. Furner. 2002. Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology 36*: 3-72.

Cheng, G. 1996. *Software Integration in a Problem-Solving Environment*. Retrieved January 2003, from http://www.npac.syr.edu/users/gcheng/homepage/thesis/node5.html

Clever Project. 1999. *Hypersearching the Web*. Retrieved March 2000, from http://www.sciam.com/1999/0699issue/0699raghavan.html

Crane, D. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: University of Chicago Press.

Crawford, S. 1971. Informal communication among scientists in sleep research. *Journal of the American Society for Information Science 22*: 301-310.

Gallopoulos, E., E. Houstis, and J.R. Rice. 1994. Problem-solving environments for computational science. *IEEE Computational Science and Engineering 1*(2): 11-23.

Garvey, W.D. 1979. *Communication: The essence of science*. New York: Pergamon.

ISI. 2000. *The ISI Database: the journal selection process*. Retrieved September 29, 2000, from http://www.isinet.com/isi/hot/essays/199701.

Lawrence, S., C. L. Giles, and K. Bollacker. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer 32*(6): 67-71.

Lin, X., H. D. White, and J. Buzydlowski. 2001. AuthorLink: instant author co-citation mapping for online searching. National Online Meeting, 15-17 May, in New York.

MacRoberts, M. H. and B. R. MacRoberts. 1989. Problems of citation analysis: a critical review. *Journal of the American Society for Information Science 40*: 342-349.

McCain, K. W. 1990. Mapping authors in intellectual space: a technical overview. *Journal of the American Society for Information Science 41*: 433-443.

Merton, R. K. 1942. Science and technology in a democratic order. *Journal of Legal and Political Sociology 1*: 115-126.

Narin, F. 1976. *Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons.

Pritchard, A. 1969. Statistical bibliography or bibliometrics? *Journal of Documentation 25*: 348-349.

Small, H. and B. C. Griffith. 1974. The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies* 4: 17-40.

Smith, L. C. 1981. Citation analysis. *Library Trends 30*: 83-106.

Walker, D. W. and O. F. Rana 2003. *Scientific problem solving*. Retrieved January 2003, from http://www.nacse.org/HPjava/walker/walker.pdf

White, H. D. 1990. Author co-citation analysis: Overview and defense. In *Scholarly communication and bibliometrics*, edited by C. L. Borgman (Newbury Park, CA: Sage): 84-106.

Zhao, D. 2003. A comparative citation analysis study of Web-based and print journal-based scholarly communication in the XML research field. Ph.D. diss., Florida State University.