Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 67 (2009) 297–312
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Designing for e-science: Requirements gathering for collaboration in CiteSeer

Umer Farooq*, Craig H. Ganoe, John M. Carroll, C. Lee Giles

College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA

Available online 17 October 2007

Abstract

It is unclear if and how collaboratories have enhanced distributed scientific collaboration. Furthermore, little is known in the way of design strategies to support such collaboration. This paper presents findings from an investigation into requirements for collaboration in e-science in the context of CiteSeer, a search engine and digital library of research literature in the computer and information science disciplines. Based on a survey and follow-up interviews with CiteSeer users, we present four novel implications for designing the CiteSeer collaboratory. First, visualize query-based social networks to identify scholarly communities of interest. Second, provide online collaborative tool support for upstream stages of scientific collaboration. Third, support activity awareness for staying cognizant of online scientific activities. Fourth, use notification systems to convey scientific activity awareness. We discuss how these implications can broadly enhance e-science usability for collaboratory infrastructures based on digital libraries.

Keywords: E-social science; Digital libraries; Scientific communities of practice; Collaboratories

1. Introduction

Digital libraries, such as ACM and IEEE, are online repositories that facilitate the first steps of scientific discovery through search and retrieval of intellectual resources. However, these digital libraries stop short of enabling "true" scientific collaboration online in the sense of providing *collaboratory* environments. Collaboratories—networked collaboration laboratories—support the social and collaborative endeavors of distributed scientists working together online. Through collaboratories, scientists can share key intellectual resources that allow colleagues located anywhere to access, view, manipulate, and have discussions about these artifacts (Kouzes et al., 1996; Finholt and Olson, 1997).

Our premise is that direct collaboration between peers in a scientific community around existing digital library resources can lead to more meaningful and long-term collaborative endeavors and scientific outcomes. The challenge we are undertaking is to enhance existing digital libraries as collaboratories for supporting distributed scientific collaboration.

In this paper, we report on requirements from our research investigation to design a collaboratory around an existing, large-scale digital library of scientific literature in computing, namely CiteSeer (http://citeseer.ist.psu.edu/). As part of our involvement in requirements gathering for the CiteSeer collaboratory, we conducted a survey and follow-up interviews with CiteSeer users. Based on data collection and analysis, we have identified four implications for design to support distributed scientific collaboration. First, visualize query-based social networks to identify scholarly communities of interest. Second, provide online collaborative tool support for upstream stages of scientific collaboration. Third, support activity awareness for staying cognizant of online scientific activities. Fourth, use notification systems to convey scientific activity awareness. These implications are novel in that they extend current findings in HCI (Human–Computer Interaction) and CSCW (Computer-Supported Cooperative Work). We discuss these implications broadly for enhancing e-science

*Corresponding author.
E-mail addresses: ufarooq@ist.psu.edu (U. Farooq), cganoe@ist.psu.edu (C.H. Ganoe), jcarroll@ist.psu.edu (J.M. Carroll), giles@ist.psu.edu (C.L. Giles).

1071-5819/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2007.10.005

usability as it relates to collaboratories based on digital libraries.

2. Motivation

Scientific communities have traditionally formed around key intellectual resources such as collections of books, or special equipment such as cyclotrons (Wellman, 1999). In the past, one of the greatest obstacles to the formation and sustained vitality of scientific communities was the fact that members had to be co-located with their shared resources and with one another.

Today, face-to-face scientific collaboration is increasingly being augmented by online interactions. This wave of online scientific collaboration has been dubbed *e-science* or *e-research*, referring to distributed and large-scale scientific collaboration enabled by Internet technologies. Through such environments, advanced computational, collaborative, data acquisition services are available to scientists and scholars in all disciplines through high-performance networks.

E-science initiatives have been taxonomized into four categories (David, 2006): community-centric, data-centric, computation-centric, and interaction-centric. Community-centric initiatives aim to bring researchers together for synchronous (e.g., chat rooms) and asynchronous interactions (e.g., discussion forums). Data-centric initiatives are concerned with storage, management, and mining of data collected from sensors, experiments, and researchers (e.g., annotating shared data). Computation-centric initiatives seek to provide high-performance computing (e.g., grid clusters). Interaction-centric initiatives enable real-time interactions for decision-making, visualization, or control of instruments where responsiveness is a priority (e.g., controlling a one-of-a-kind telescope).

Community-centric e-science initiatives have been scarce. For example, of the 23 pilot projects funded under the UK e-science core program, 16 are data-centric and only one is community-centric (David, 2006). There is a more uniform distribution across the taxonomy of projects funded in the United States by the National Science Foundation during the late 1980s and early 1990s, although most community-centric collaboratories were organized around a single research project (David and Spence, 2003, p. 70). Lack of community-centric e-science initiatives does not imply that community building has faded altogether, but it no longer appears to be so central an objective for the developers of collaborative e-science infrastructures (David and Spence, 2003, p. 68). We take this gap as a challenge to develop a collaborative e-science infrastructure to foster community building and collaboration among geographically distributed scientists.

Often, the GRID is referred to as a possible platform for global collaboration, communication, and e-science (Hinde and Wilcock, 2002). The premise is that the use of a large, distributed computational fabric based on GRID technologies, can be used to provide a scalable communication platform and media processing fabric to support a high-quality communication and collaboration environment. The foundations for the use of the GRID as middleware to provide global collaborative services in the areas of science and engineering were laid down by Ian Foster, Carl Kesselman, and Stephen Tuecke. In their Globus project, they developed parts of the prototype of the open-source GRID toolkit (Foster and Kesselman, 1998).

E-science initiatives in the UK have led to concrete results in terms of creating professional venues for sharing and disseminating developments in this key area. For example, the main e-science conference, All Hands Meeting (AHM: http://www.allhands.org.uk/), has been held four times. Many workshops to further the understanding of collaborative e-science infrastructures have been established. For example, UK's Joint Information Systems Committee (JISC) held two workshops on the topic of "Building collaborative e-research environments" in 2004 (http://www.jisc.ac.uk/index.cfm?name=event_eresearch/). Web repositories of e-science developments have been active in assembling an online forum of publications and discussions (e.g., UK Research Councils: http://www.rcuk.ac.uk/escience/).

In addition to the UK e-science programs, attention has been recently focused toward Australian e-science initiatives. In a technical report by The Australian Academy of Technological Sciences and Engineering (ATSE), it was asserted that UK e-science initiatives have focused on big science areas such as particle physics and astronomy whereas Australian initiatives have addressed their needs in the areas such as biotechnology, health, and mineral processing (Sargent, 2004). As a result of this report, many opportunities for collaborative participation were identified between the UK and Australian e-science programs.

The United States, through its National Science Foundation, has been involved in e-science initiatives, referring to them as collaboratories to mean a "center without walls, in which the nation's researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resource, and accessing information in digital libraries" (Wulf, 1993). The challenges and opportunities in creating collaboratories and their interfaces relate directly to many aspects of HCI and CSCW. As a result of collaboratory development and HCI/CSCW research converging, a special issue of ACM Interactions was published in 1998, comprising four key articles that offered an in-depth look at collaboratories. An online list of collaboratories is also available (Science of Collaboratories, 2006).

Scientific research collaboration is increasingly being seen as critically dependent upon effective access to, and sharing of digital research data (David, 2006). The original justification for collaboratories, for example, was a matter of resource access and logistics than of supporting and enhancing collaboration (Farooq et al., 2005; Farooq, 2006). It is equally critical, though, to be able to access

one's collaborators anywhere and anytime. Collaboration supports the extension of e-science to online communities that transcend current organizational and geographical boundaries (De Roure et al., 2001).

Little research has been done in fostering distributed scientific collaboration directly between peers in communities. Community-centric e-science projects have lacked the vision and depth in this regard. It is not as simple as supporting interaction between community members through synchronous and asynchronous information exchanges mediated by technologies such as chat rooms, discussion forums, and video conferencing, as some may have initially thought. What if members in a scientific community do not want to collaborate? Even if they do, with whom do they want to collaborate? How do they want to collaborate? What kinds of socio-technical issues will they face in collaboration and how will they be addressed?

All these aforementioned issues are critical because scientific communities do not exchange mere information. They exchange and also create knowledge, both tacit and explicit. Scientific communities, like other communities, are dynamic—they grow, evolve, and change over time. Members of a scientific community are held together by their common intellectual resources—the ideas they produce, the similar literature they refer to, and the papers they publish to address common problems.

In this paper, we seek to answer some of the open-ended questions and issues listed above. The context of our investigation is most directly related to HCI and CSCW issues in collaboratories. In this sense, we are adopting the perspective that e-science is about developing technologies that give rise to "virtual organizations" (Foster and Kesselman, 1998) for the purposes of sharing information and knowledge, collaboration among peers, sustaining scholarly communities, and so forth.

3. Related work

Little is known in the way of design strategies to support distributed scientific collaboration because only a few collaboratories have been evaluated from this angle (e.g., Sonnenwald et al., 2003; Olson et al., 2008), resulting in just a handful of basic design issues and heuristics related to general collaborative experiences in collaboratories (e.g., Finholt, 2002; Olson et al., 2008). Below, we review major findings from prior work in order to ground our contribution.

In 2002, Finholt (2002) wrote a retrospective article in which he outlined a number of design principles related to collaboratory development. More recently, Olson et al. (2008) proposed a work-in-progress theory of remote collaboration based largely on their experience with collaboratory development. Other classical studies, such as by Star and Ruhleder (1994, 1996), have also analyzed distributed work in collaboratories, but these two articles represent the state-of-the-art in designing collaboratories to support scientific collaboration. We have codified the design principles from these sources below, also summarized in Table 1.

A classical finding in CSCW is that co-location is essential for tightly coupled tasks. This is one of the reasons that virtual collaboration is so difficult to achieve. Based on laboratory experiments and empirical investigations, Olson and Olson (2001) conclude that for tasks depending on frequent interaction and feedback among collaborators, contemporary computer mediated communication technologies such as email and audio/video conferencing do not provide adequate substitution for face-to-face interactions.

Even if enhanced technologies are identified to compensate for lack of co-location in virtual collaboration, there are still barriers to successful online interactions. Olson et al. (2000) assert that collaboration readiness and collaboration technology readiness are two such barriers. Collaboration readiness refers to the extent of the motivation and desire to work with each other. Even if co-workers are open to collaboration, the underlying technology infrastructure and availability of local technology expertise may be insufficient. Such collaboration technology readiness is also essential for successful virtual collaboration.

Based on a review of studies of scientists and engineers at work, Finholt (2002) claims that scientists who are remote from communities of elite and active researchers are at a disadvantage in terms of initiating contact with leading investigators that can lead to deeper collaboration. Therefore, improved access to important but scarce instruments and data, combined with communication among researchers, can reduce this status misalignment and facilitate increased involvement by non-elite scientists in cutting edge research.

Clark and Brennan (1991) have identified that lack of common ground (e.g., mutual knowledge) and trust can hinder collaboration. Olson et al. (2008) have corroborated these findings in their investigations with collaboratory interactions among scientists, further emphasizing the role of history and shared vocabulary among collaborators as an essential facet of successful virtual collaboration.

Olson et al. (2008) note in their collaboratory studies that providing support for processes of management, planning, and decision-making is important. When there are multiple institutions involved and/or different departments (disciplines) within the same university, it is important to know who is serving in what role. Those projects left loose suffer when the participants' directions begin to diverge; they have not assigned someone to take leadership to get the group back on track, and have not bought in to that person having that authority. The larger the collaboration (such as in a collaboratory), the more important these elements (like management) become (Cummings and Kiesler, 2005).

Whereas the above-mentioned design principles serve as useful heuristics for developing collaboratory technology, they are far too general to indicate specific design tools to

Table 1
Design principles from prior work in context of collaboratories

Summary	Source
Co-location is required for tightly coupled tasks.	Olson and Olson (2001)
Collaboration "readiness" and "technology readiness" are essential factors for success.	Olson et al. (2000)
Status misalignment can hamper communication between scientists (e.g., elite vs. non-elite).	Finholt (2002)
Lack of common ground (e.g., mutual knowledge) and trust can hinder collaboration.	Olson et al. (2008)
Management, planning, and decision-making are important processes to provide support for.	Olson et al. (2008)

support the unique social nuances of distributed scientific collaboration. According to Dourish (2001), design principles are in fact not design recommendations, rules, or guidelines. Instead, these principles observe or comment upon general characteristics of desirable design features. In a sense, they act as heuristics that embody design tradeoffs, which help the designer formulate specific design tools and their features.

Our contribution in this paper differs in at least two ways from prior collaboratory investigations. First, we are starting from a requirements gathering phase whereas the above design principles were mostly founded on post-implementation experience of the collaboratories. Therefore, we see our contribution as complementing existing empirical findings through a systematic process of software development.

Second, we are not proposing design principles in the sense of how Dourish (2001) defines them, but design *recommendations* that have direct design consequences; that is, design can be operationalized through implementation. Of course, it is our hope that implementation of our design recommendations will lead to further refinement and eventually the abstraction of general design principles from empirical evaluations (similar to the ones listed in Table 1).

4. Background of CiteSeer

Our study context is CiteSeer (Giles et al., 1998): a search engine and digital library of research literature in the computer and information science (CIS) disciplines that is a free public resource providing access to the full-text of nearly 700,000 academic science papers, and over 10 million citations. CiteSeer currently receives over approximately half a million hits a day and is accessed by 150 countries and 200,000 unique machines monthly. CiteSeer was created by Kurt Bollacker, Lee Giles, and Steve Lawrence in 1997–1998 at NEC Laboratories to provide a comprehensive and accessible digital library and search engine for the CIS research community. The query "CiteSeer" returns millions of unique documents from the popular Internet search engines Google and Yahoo and is widely indexed by both. CiteSeer is frequently cited as a search service that has greatly improved communication and progress in computer science research. It is currently hosted and maintained by the College of Information Sciences and Technology at The Pennsylvania State University.

It is traditional practice in the CIS community to make research documents available at the time they are first written through technical reports series managed by various laboratories and academic departments. More recently, this practice has been transferred to the World Wide Web (Goodrum et al., 2001). CiteSeer actively and automatically harvests these documents and automatically builds searchable and indexable collections, promoting creative scientific discovery and reuse within the CIS community. Even though search engines such as Google actively index CiteSeer, users come to the CiteSeer engine for unique information such as citation counts and domain dependent citation links not provided by Google or Google Scholar. CiteSeer is a full text search engine with an interface that permits search by document, numbers of citations, or by fielded searching, not currently possible on general-purpose web search engines. Fig. 1 shows a screenshot of CiteSeer's interface that captures the primary information that the scholarly digital library currently provides.

5. Methods

The broader goal of our research investigation is to enhance CiteSeer as a collaboratory in order to support distributed scientific collaboration among scholars around their intellectual resources in the domain of computer and information science. The idea is to provide collaboratory services as a value-added layer on top of CiteSeer's information search and retrieval interface. As a first step, we wanted to gain insight into the kinds of activities CiteSeer users would like to collaborate on and possible socio-technical issues during such collaboration. We conducted an online survey study with CiteSeer users to achieve this goal.

5.1. Recruitment and participants

The survey was made available on CiteSeer's web site. Thus, this is an opportunity sample. Participants were CiteSeer users willing to take the online survey. The survey link was placed in multiple locations on CiteSeer's web site, and read, "Help us improve CiteSeer. Take a survey". Clicking on the survey link would direct users to an informed consent web page and upon acceptance of the consent form, users would be redirected to the survey questions. No compensation, financial or otherwise, was provided to survey responders.

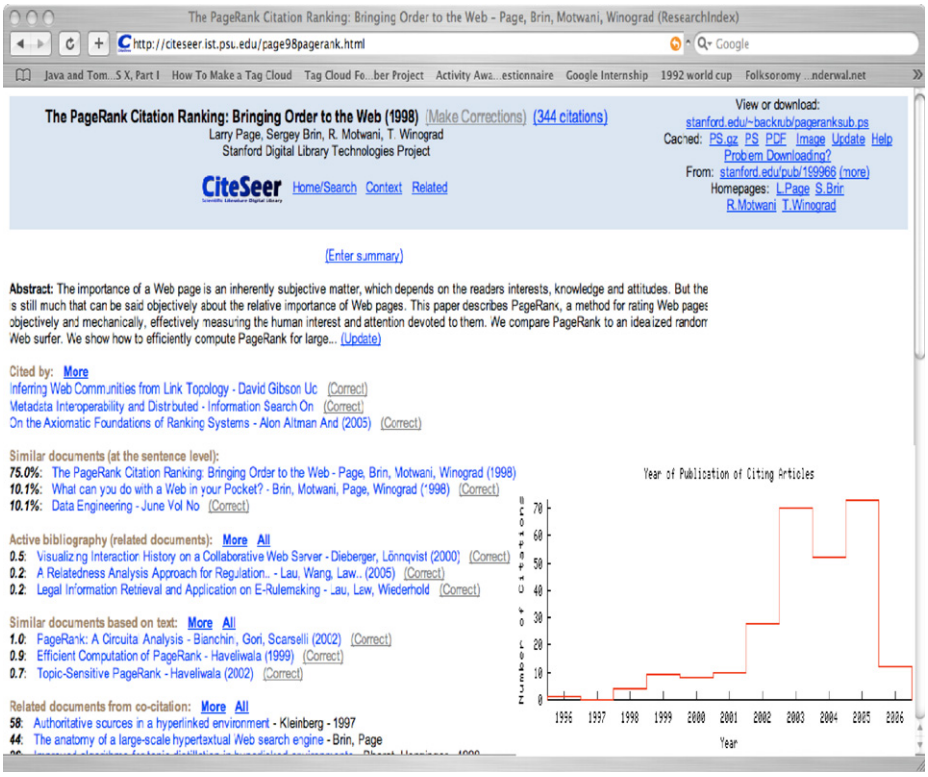

Fig. 1. Screenshot of CiteSeer's current functionality.

The survey link was placed in multiple locations on CiteSeer's web site to increase the probability of the number of participants responding to the survey. The primary location of the survey link was on CiteSeer's home page (http://citeseer.ist.psu.edu/). However, because CiteSeer's resources are indexed by other search engines (e.g., Google), people can be pointed to CiteSeer's resources directly, bypassing the home page. To accommodate such users, the survey link was also placed on the results web page (when CiteSeer is queried) and details web page (when a specific resource is selected). In addition, the survey link was also placed on all web pages that required CiteSeer users to enter information such as correcting a paper's citation. Examples of such web pages included the update web page (for updating information related to a paper), comments web page (for commenting on an article), and feedback web page (for providing feedback to CiteSeer administrators).

Our results in this paper are based on the administered survey for 2 weeks (November 17–30, 2005). The number of participants who responded to the survey was 301. Some respondents skipped one or more questions.

5.2. Survey design

We designed the survey asking 29 questions (12 of which were multi-part) organized into four broad sections:

- *Professional interaction*. Seven questions (five of which were multi-part) related to how often CiteSeer users collaborated face-to-face and remotely, how they would like to collaborate with other CiteSeer users, and what issues they might face in online collaboration.
- *CiteSeer use*. Seven questions (two of which were multi-part) related to how often CiteSeer users used the search engine, the nature of CiteSeer queries, and whether or not the use of CiteSeer led to collaboration.
- *Comparison of search engines with CiteSeer*. Six questions (five of which were multi-part) related to the use of CiteSeer with other academic search engines such as ACM Digital Library, IEEE Xplore, and Google Scholar.
- *Background information*. Nine questions related to demographics of CiteSeer users.

The questions were predominantly a mix of selection among pre-defined categories (e.g., age ranges, frequency of CiteSeer use) and ratings on seven-point Likert scales (e.g., frequency of use for a specific CiteSeer feature on a scale of "Never" to "Very often"); few free-text opportunities were provided (e.g., academic background). Based on pilot testing, the survey required approximately 10–15 min to complete.

5.3. Data collection and analysis

5.3.1. Quantitative data

Most of the survey questions solicited numerical responses. Analysis of this quantitative data was performed using the statistical package SPSS. A variety of statistical

analyses were performed, such as analysis of variance (ANOVA) and correlation analysis.

Because we included multi-part questions in the survey, it was important to check the reliability of the scales, that is, to make sure all the items in a question were measuring the same underlying construct. The scales on the multi-part questions had good internal consistency, with all Cronbach α coefficients reported above 0.7.

5.3.2. Qualitative follow-up data

The last survey question asked participants if they were willing to be interviewed via email or phone. We contacted 66 participants who were willing to be interviewed via email and got responses from 22 participants. The second to last question in the survey asked for any type of feedback from participants (e.g., related to CiteSeer or the survey); 94 participants responded to this question. We analyzed these responses as well for any feedback related to the interview questions.

We asked the following four questions in the email interview: (1) Which criteria would you find most important for collaborating with CiteSeer users, and why? (2) Which online collaborative activities would be most valuable to you, and why? (3) Which activities would you like to stay most aware of, and why? (4) What would be the best way for you to stay aware of these activities, and why?

6. Survey results

Of the responses we received, 42% were graduate students. Males (89%) outnumbered females. More than half the respondents (52%) were in the age range of 21–30 years.

Forty-two percent of the respondents had a master's degree. The sample as a whole was relatively highly educated, with 32% having a doctorate degree. Because CiteSeer is a digital library primarily for the CIS disciplines, it was consistent that 79% of the respondents had a computer science background.

The survey respondents represented a relatively core group of CiteSeer users. Their mean (M) use of CiteSeer was reported as 3.7 years (S.D. = 1.7). Almost half (45%) claimed to have downloaded more than 100 papers from CiteSeer. 40% said they use CiteSeer once or twice per week.

We present the survey results under the following three themes: (1) *Potential collaborators*; (2) *Online collaborative activities*; and (3) *Awareness issues*.

6.1. Potential collaborators

We wanted to understand with whom CiteSeer users would collaborate online. Participants were asked to rate how often they would like to interact remotely with others on a scale of 1 (Never) to 7 (Very often) based on six items: (1) Who are looking for similar types of papers as I am; (2) Who read my papers; (3) Whose papers I read; (4) Who cite my papers; (5) Whom I cite in my papers; (6) Who cite similar papers as I do.

The six items were rated relatively high with all means above 4 (Sometimes): 4.76, 5.03, 5.10, 5.00, 4.97, and 4.65, respectively. Because the quantitative data is inconclusive, it is unclear which of the six criteria will be most useful to match potential collaborators.

Qualitative data prioritizes some of these criteria. For example, people are likely to collaborate with those who look for similar papers and read each other's papers. Reading similar papers is an indicator of people working in the same area, as one respondent suggests:

> Important criteria: users who are reading the same and similar papers as me. Since we are reading the same papers, we are working in the exact same sub-area.

It seems plausible that someone who looks for similar papers as another person also cites similar papers. In this case, potential collaborators can share common ideas that focus on the papers they look for or cite. One interview respondent expressed this view:

> [I want to] collaborate with CiteSeer users who are looking for similar papers as [me] and who cite similar papers as [I do] … the reason is I can save more time to find a good paper worth reading and can touch more ideas in my research area by collaboration.

We ran a correlation analysis between the question items "Who look for similar papers" and "Who cite similar papers". The Pearson product–moment correlation coefficient was 0.66 ($N = 217$, $p < 0.01$), which corroborates the above qualitative feedback.

A concern in matching people based on readings or citations is the use of personal, sensitive information. Surprisingly, no one indicated that using personal information would be an issue. On the contrary, one interview respondent suggested that people's web sites could be used to identify potential collaborators:

> [For] connecting users with common interests … focus on researchers' home pages, because almost everyone I have seen from academia gives a links page …

One interview respondent provided an insight into how matching potential collaborators can also facilitate opportunistic collaboration outside of one's research area and expertise:

> [An] important aspect to collaboration is to facilitate 'serendipitous' interaction. As it is said, it's not what you don't know, it's what you don't know that you don't know. This is closely related to the discovery of cross domain knowledge and expertise.

The quantitative part of the survey did not probe users about the representation of social matching. As indicated by many survey respondents, social networks are

appropriate for depicting meaningful social structures in CiteSeer:

> I think it would be great if I could get a CiteSeer page with a 'network' diagram … and 'related' strong links and more remote links clearly shown.

6.2. Online collaborative activities

We wanted to know what kinds of collaborative activities CiteSeer users would like supported. Participants were asked to rate how often they currently interact with others on a scale of 1 (Never) to 7 (Very often) based on four items: (1) Strengthen social connections; (2) Brainstorm new ideas; (3) Plan joint projects; (4) Write joint papers.

In general, respondents rated all items moderately high with all means above 4 (Sometimes): 4.28, 4.71, 4.32, and 4.06, respectively. Participants were also asked how difficult they would find these activities to achieve remotely on a scale of 1 (Very easy) to 7 (Very difficult). Responses indicate that CiteSeer users found these distributed collaborative activities to be on the difficult side of neutral (4), with respective means as 4.40, 4.36, 4.53, and 4.47.

One interpretation of these results is that CiteSeer users moderately engage in these types of collaborative activities. However, remote collaboration is perceived as somewhat difficult. Qualitative data elaborates on the kinds of online activities that CiteSeer users would like supported and gives reasons for not supporting other activities that they perceive as difficult.

Overwhelmingly, online discussions forums were the most popular type of distributed collaborative activity, as indicated by one of many respondents:

> I'd love to participate in forums or discussions about my field, to see what is going on, and what other people think.

Discussions can also be a valuable source for new ideas. The following interview respondent indicated the fact that discussions can enable brainstorming:

> [I would be interested in] brainstorming new ideas related to online discussions.

Given that, in general, science is a social enterprise, one would expect CiteSeer users to want online support for collaborative activities such as joint planning and shared writing endeavors. However, according to our interview respondents, they are not inclined to use such collaborative features. One interview respondent said:

> Writing new papers and planning projects don't seem like activities people would actually do through a science portal.

This respondent's view was corroborated by others who thought that current ways (e.g., email) of achieving such joint endeavors would suffice:

> I think the online discussions and brainstorming could be useful. For paper writing and project planning, I'd imagine that the team would be cohesive and we'd just use email or a wiki to coordinate.

We wanted to understand how collaboration for CiteSeer users differs between face-to-face and remote settings. In the survey, participants were asked to rate how often they interact professionally with others within (face-to-face) and outside (remote) of their lab or institution on a scale of 1 (Never) to 7 (Very often). For professional interaction within their lab or institution, 42% ($N = 293$) of the respondents chose a rating of 7 (Very often). The mean response was 5.59 ($N = 293$), indicating that face-to-face interaction within lab or institutional vicinities is a primary method for collaboration. For professional interaction outside of their lab or institution, 36% ($N = 287$) of the respondents chose a rating of 4 (Sometimes). The mean response was 4.39 ($N = 287$), indicating that remote interaction outside of lab or institutional vicinities is a moderate method for collaboration.

A paired-samples t-test was conducted to evaluate for a statistically significant difference in the mean scores between the level of face-to-face and remote collaboration. A statistically significant difference was reported ($t(285) = 13.54$, $p < 0.0005$). The η^2 statistic (0.39) indicated a large effect size. Our qualitative data sheds light on this difference.

Trust and privacy are two factors that can distributed collaboration. One respondent said:

> Collaboration is based on mutual trust, and it cannot be gained easily via an Internet site. Also, the question of privacy comes to my mind—one would not be willing to share his preliminary ideas to an unknown audience.

Establishing trust and privacy becomes more difficult when potentially valuable ideas, which form the basis of scientific discovery, cannot be shared due to institutional constraints, or are shared and unethically misused. For example, legal issues can hinder distributed collaboration, as indicated by the following interview respondent:

> Some people will, no doubt, wish to be 'silent participants' [in online collaboration] due to legal intellectual property issues.

6.3. Awareness issues

We wanted to understand awareness issues in online collaboration. Participants were asked to rate their level of agreement on how difficult they find it to stay aware of CiteSeer resources on a scale of 1 (Strongly disagree) to 7 (Strongly agree) based on four items: (1) Recent papers published in my area; (2) Who reads my papers; (3) New colleagues who are working in my area; (4) Who cites my papers.

Results suggest that staying aware was generally difficult as at least 50% of all respondents rated all items toward the

agreement side of the scale. One-way within-subjects ANOVA was conducted with the awareness resources as the independent variable with four levels (the response items) and level of difficulty (rating from 1 to 7) as the dependent variable.

The Levene test was significant at 0.001, so the assumption of homogeneity of variance was violated. Therefore, both Brown–Forsythe and Welch F-ratios are reported. The ANOVA was significant, with $F(3, 594.44) = 22.68$ ($p < 0.0005$) and $F(3, 1057.04) = 22.08$ ($p < 0.0005$), respectively. We computed a contrast test between the first item (recent papers published in my area) and the other three items combined. Results indicate that the first item was rated significantly lower, with $F(1, 472.07) = 37.27$ ($p < 0.0005$). Thus, CiteSeer users find it less difficult to stay aware of recently published papers in their area, perhaps because this is done traditionally (through subscriptions to journals and conference attendance).

Although our quantitative questions only asked about the difficulty in staying aware, qualitative data suggests that awareness of CiteSeer resources and activities of CiteSeer users around those resources is important. An interview respondent said:

> [The most interesting awareness feature is] providing statistics on your own papers (readers, citations).

Staying aware of new colleagues in one's research area is also important to keep abreast of potential collaborators, their activities, and their research focus. An interview respondent said:

> I'd like to know who has started a new discussion thread related to my area of interest, because I want to be aware what is going on outside my lab, and what other researchers are thinking or focusing on.

Qualitative data also suggests that mining historical activities in CiteSeer to provide influence patterns and impact assessment of intellectual resources can enrich awareness information. An interview respondent indicated the relevance of history for awareness and how it can also inform future impact of a discipline:

> It's always important to be aware of new research efforts starting up that are synergistic or disruptive relative to your own. You might consider online 'analytics' that give people some idea of where activity is centered and where it is going … It could tell you if, for example, interest in a discipline is 'dying down' or 'ramping up'.

In addition to historical information, supporting awareness of current activities is important in order to stay cognizant of up-to-date information. For instance, CiteSeer users want to be notified when a specific event has taken place, as indicated below:

> I would find it more important to know when a paper was entered into CiteSeer that cited one of my papers; that would be a strong signal that I might have interest in it.

An important facet of awareness is how it will be conveyed. Many interview respondents indicated the usefulness of Really Simple Syndication (RSS) feeds:

> [I] definitely [want] RSS: it isn't intrusive (I get information when I want), information can be easily [and] automatically processed, [and] I can get information in whatever way I want (as emails, in my aggregator, in my browser, …).

In addition to how awareness information can be conveyed, respondents indicated different types of information they would like to stay aware of. One respondent wanted to know about "hot topics" (implying popular topics) being discussed in forums. In another example, a respondent was interested in papers for a specified area of interest (e.g., using keywords) or those that cite his/her work:

> Features that would be useful are alerts when new articles are posted that either contain keywords or cite work I am interested in to keep abreast of what's new in my field.

Even though there are traditional ways of staying aware of new papers, using features to refine such awareness (e.g., through keywords) seems desirable.

7. Implications for design

Several of our results suggest specific strategies to support distributed scientific collaboration. The four implications for design are the following: (1) Visualize query-based social networks to identify scholarly communities of interest. (2) Provide online collaborative tool support for upstream stages of scientific collaboration. (3) Support activity awareness to stay cognizant of online, asynchronous, and long-term scientific activities. (4) Use notification systems to convey scientific activity awareness peripheral to users' primary task.

The implications are motivated by design rationale based on survey results and related HCI/CSCW literature. Design envisionment scenarios, conceptual schemas, and prototype screenshots are used to illustrate the implications for design.

7.1. Visualize query-based social networks

In regard to matching potential collaborators, survey results support existing claims. Literature from social psychology asserts that people are attracted to "similar others" (Terveen and McDonald, 2005, p. 416), with similarity in interests being one facet of this. In the area of knowledge management, recommender systems have looked at identifying individuals who have expertise in

one's area of interest (see McDonald, 2001 for a comprehensive review).

In CiteSeer, identifying users with similar interests can be based on multiple criteria, such as mutual reading of papers, citations, and similar search behavior. Similar search behavior seems to be a feasible candidate among these choices for at least three reasons.

First, CiteSeer can easily keep track of users' search behavior by storing and mining a history of user queries (users would have to provide consent for CiteSeer to track their behavior). CiteSeer queries—typically, noun phrases such as "user-centered design"—essentially filter the space of available resources into specialized views. These views can be thought of as research investigations, research areas, or even sub-disciplines. Many queries are in effect reused in the sense that someone else entered that query, or one like it, before. Comparing these queries with similarity measures can provide social matching heuristics for users.

Second, search queries are universal. For example, social matching based on citations may not apply to all users as everyone would not have a critical mass of cited papers (e.g., new researchers).

Third, queries accurately convey first-hand information about user interests. Queries that cumulate over time related to the same topic can indicate a strong interest in that topic. Of course, two users submitting similar queries do not necessarily want to collaborate, but the chance that collaboration would be attractive at some level is more likely than individuals with totally different interests.

Scholarly communities and sub-communities can form around queries, just as they have traditionally formed around shared resources. Providing a virtual place for scientists with common query interests to share perspectives, related and updated information and links, and so forth would enrich these queries for everyone, and help scholarship and scholarly communities of interest or practice to form and develop (Wenger et al., 2002).

These scholarly communities could be codified through social network analysis where shared queries are the primary basis for links among persons in the network. Query-based social networks would connect persons more or less directly, depending on how many queries they shared, and how they were connected to others in the network. We might expect interesting community phenomena to emerge from such networks. For example, the network could foster scientific collaboration, not just between members within a particular scientific group, but also between *weak ties* (Granovetter, 1973), scholars principally belonging to different groups who are connected through others. This can help CiteSeer users to identify new colleagues and potential collaborators more easily. One issue to consider is the potential mismatch between the social networks that CiteSeer provides and the perception of users on what their social networks should look like. Providing transparency, such as by explaining why the social network is constructed as it is, may mitigate this issue (Herlocker et al., 2000).

Social structures can also be used to discover and reinforce cross-community *bridges*. Bridges are, at the most basic level, members of two or more distinct community organizations (Kavanaugh et al., 2005). In a scientific community, bridges are researchers who are part of two or more research communities (e.g., HCI and IS: Information Systems). Through query-based social networks, scientists can opportunistically explore nodes and edges beyond their immediate task goals, and learn about bridges and their expertise that complement their own research area. Scientists who expand the edges of their communities in this way become more aware of activities that might influence their own work. This perspective aligns well with our survey results that indicated the advantages of serendipitous collaboration.

An issue in social matching systems is the use of personal information. Personal information is critical for matching people. Terveen and McDonald (2005) claim that social matching systems need to use—and users will be willing to supply—relatively personal sensitive information to effectively match people. It is worthwhile to note that while Terveen and McDonald meant personal sensitive information to imply age, music taste, hobbies, and so forth, we are construing such information for scientific communities as user queries. We anticipate few problems in getting scientists to allow use of their queries (anonymous to other users) for system-level social matching, given that evidence suggests that people will be willing to share more personal sensitive information, as per Terveen and McDonald's claim. Survey results mildly indicate that users would be willing to provide such information, such as their personal web sites.

Example. Consider the conceptual schema of a possible query-based social network in Fig. 2. Joe, Ike, and Bo are strongly connected to each other in a social network based on their shared query "algorithms and databases". Bo, Zaz, and Ali are also strongly connected to each other based on their shared query "bio-informatics". Because Bo shares both queries, he is a bridge. Thus, Joe and Ike are connected to Zaz and Ali through Bo. Joe is surprised to find out that his most recent algorithm to optimize data storage in a relational database is being used in bioinformatics to efficiently store, access, and visualize patient care records. Joe is excited to contact Zaz to further cross-community collaboration and refine his algorithm.

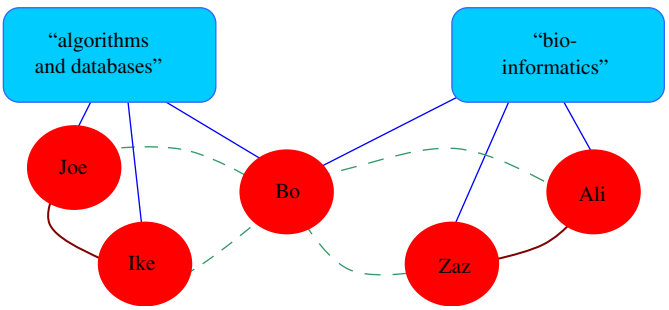

Fig. 2. Conceptual schema of a social network.

7.2. Support upstream stages of collaboration

Survey results suggest that CiteSeer users would welcome opportunities that support open-ended and idea-generation activities. Contrary to focused activities, we characterize such opportunities as upstream stages of scientific collaboration. This refers to early, divergent stages of scientific discovery in contrast to final, convergent stages.

While asynchronous discussion forums are ideal for open-ended and divergent technical discussions, they are often not flexible or interactive enough to support finer-grain collaborations like joint authoring (Kelly et al., 2002). Thus, it was our intention for CiteSeer users to be able to create collaborative spaces for more synchronous, sustained, and convergent collaborative interactions for developing intellectual products such as research papers and proposals. However, survey results strongly suggested against such focused tool support.

In addition to issues such as trust and privacy, results indicated that users did not want support for such focused collaborative activities because they already had existing ways of engaging in such endeavors. Some respondents suggested that face-to-face interactions and email are sufficient to achieve focused activities. Hollan and Stornetta (1992) assert that face-to-face interactions cannot be replaced by any other collaboration channel, and therefore, the goal of developing tools for distributed interaction should be to identify needs that are not met in physical proximity. Olson and Olson (2001) also suggest that co-location is still essential for some collaboration, especially for tightly coupled and focused activities that demand frequent interaction and feedback among participants.

The design rationale for supporting upstream stages of scientific collaboration stems from at least two reasons. First, lack of common ground, trust, and intellectual ownership should be less important issues at the preparatory rather than concluding stages of scientific collaboration. This is because the goal during upstream stages of collaboration is to generate, share, and leverage ideas with a communal orientation. During these stages, diversity of perspectives is especially useful and important (Levine and Moreland, 2004). The benefits of collectively engaging in such collaboration are likely to outweigh its costs.

Second, supporting upstream stages of scientific collaboration represents a segue from just search and retrieval tasks of CiteSeer's resources to interacting minimally with other users. This is consistent with the existing finding that technology readiness is required for successful collaboration (Olson et al., 2000). Attempting to leapfrog steps by providing sophisticated applications (e.g., collaborative writing tools) rather than progressive interventions can produce frustration and resistance on part of the users.

Survey results provided examples of tools that could support upstream stages of scientific collaboration. Discussion-oriented tools were a popular demand. CiteSeer currently can directly present the influence network for a resource (e.g., a listing of papers that cite Grudin's paper "Groupware and social dynamics: Eight challenges for developers"), but it does not provide a textual exegesis synthesizing and interpreting that network of citations (e.g., discussions on the ideas in Grudin's paper, their influence on particular researchers, and so on). Such an exegesis could be the social construction of a scientific community. Providing an explicit medium to codify such discussions can enrich the specific resources for everyone who accesses them, and more generally can help scholarship and scientific communities develop.

In addition to discussion tools, collaborative brainstorming tools such as concept maps and white-boards are likely to support scientific discovery. It has been shown that brainstorming can increase the ability to share and generate creative ideas (Sutton and Hargadon, 1996).

Example. We have been prototyping a workspace that supports upstream stages of scientific collaboration within CiteSeer. Fig. 3 shows a screenshot of this prototype. The idea is that CiteSeer users can engage in synchronous and asynchronous brainstorming activities, such as through collaborative concept maps and threaded discussions. The timeline on the top maintains version histories of collaborative activities.

This workspace is part of a larger, integrated toolkit known as BRIDGE: Basic Resources for Integrated Distributed Group Environments (http://bridgetools.sourceforge.net) (Ganoe et al., 2003). BRIDGE clients are seamlessly integrated with browser-based Wiki-style asynchronous editing, and they also support synchronous shared editing of complex documents. For accessibility and familiarity, BRIDGE clients look and behave like a normal web site, with all content rendered as HTML and images. Simple forms of authoring are supported. Each page has an "Edit" link which supports editing and new page creation

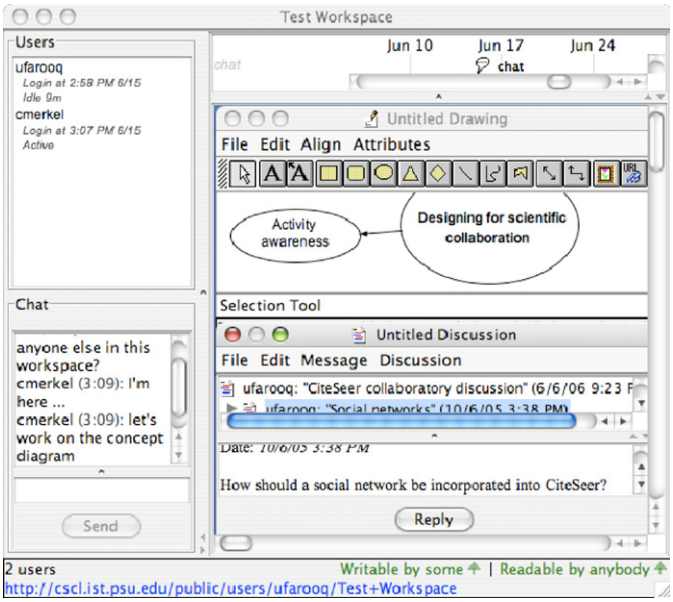

Fig. 3. Collaborative workspace prototype.

using a simple shorthand notation that requires no external authoring tools or knowledge of HTML. Each BRIDGE web page also has a "Full Editor" link that launches an interactive Java-based client. The Java client supports interactive authoring functionality that is not possible or practical using HTML-based forms, and attempts to integrate tools, some of them that facilitate creativity, supporting rich collaborative activities. In our current implementation, this includes tools for drawing (creative thinking by free associations), creating data tables, charts, and interactive maps (creative composition of artifacts and performances).

7.3. Support activity awareness

CSCW literature has highlighted the importance of awareness for successful collaboration (Dourish and Bellotti, 1992). For example, it is critical to know who else is present—social awareness (Erickson et al., 1999)—and what others are doing—workspace awareness (Gutwin and Greenberg, 1996)—in a shared workspace.

Survey results suggest that supporting awareness for CiteSeer users is an opportunity. Traditional types of CSCW awareness mechanisms are unable to adequately support the type of desired awareness features expressed by survey participants. This is because awareness in CSCW has focused more on supporting synchronous mediums of interactions over brief periods of time (Carroll et al., 2006a). Asynchronous and long-term awareness phenomena have been investigated somewhat less. Furthermore, investigating awareness specifically for scientific collaboration has not been explored before.

Survey results suggest that for CiteSeer, most activities are asynchronous and long-term. For instance, users use CiteSeer intermittently over the course of months and years, depending on when they need to access intellectual resources. This implies that although staying aware of what is going on at the present time is important, awareness of historical and future activities is key to successful collaboration.

Activity awareness (Carroll et al., 2006b) has sought out to provide such activity-based information. Activity awareness is awareness of project work that supports group performance in complex tasks over long-term endeavors directed at major goals. Activity awareness allows reflection of one's work, review of prior session histories, and analysis of future collaborative endeavors.

The design rationale for specifically using activity awareness is grounded in activity theory (see Carroll et al., 2006a for details). An activity-centered perspective emphasizes complex socially and culturally embedded endeavors that are organized in dynamic hierarchies. For example, scientific activities involve convergent and divergent thinking, development of professional relationships with peers, collaboration with others that unfolds over time, dissemination of work in the broader scientific community, and so forth.

The argument is that collaborators, such as scientists, need to be aware of one another's activity, understood in this broad sense. Carroll et al. (2006a) described a framework for activity awareness that takes the perspective of individuals operating within *communities of practice* (such as a scientific community) that emerged and are sustained through the construction of *common ground*, exchange of *social capital*, and the processes of *human development*. Such a framework is highly appropriate for supporting scientific activity awareness during distributed collaboration. This is because scientists have personal goals for contribution and reputation (human development), as they collaborate with peers (social capital) based on mutual trust and knowledge (common ground), operating in a globally distributed research environment (community of practice) (Carroll et al., 2006b).

Example. Consider the following scenario: To assess the impact of her research, Diane signs up to be alerted when any of her papers in CiteSeer are cited. She also subscribes to a service for notifying her when new papers in her research area are available. While editing her paper, Diane receives a notification that Larry Somers has just published an article in her flagship journal. She shares this article with her graduate students to discuss how their proposed experiment can build on the article's empirical results.

In this scenario, scientific activity awareness is supported through a subscription service that computes influence patterns of papers based on citations. Scientific activity awareness was also used to keep track of latest research in a community of practice (community of scholars in one's research area). Here, Diane's immediate research is affected by the publication of a recent journal article that generates social capital in her research group. Activity awareness allows one to keep abreast of such online, asynchronous scientific activities over time.

7.4. Use notification systems

Part of the challenge in supporting computer-supported awareness is knowing *how* to convey it effectively. Survey results suggest that alerting services like RSS are preferred. We refer to such awareness mechanisms as notification systems. Notification systems appear to be a reasonable mechanism to convey scientific activity awareness. Notification systems are defined as interfaces that are typically used in a divided-attention, multi-tasking situation, attempting to deliver current, valued information through a variety of platforms and modes in an efficient and effective manner (McCrickard et al., 2003). They are typically lightweight, event-triggered displays of information peripheral to a person's current task-oriented concern, for example, system status updates, email alerts, stock tickers, and chat messaging. Notification systems have been used before to support collaborative activity awareness (Carroll et al., 2003).

The design rationale for using notification systems to convey scientific activity awareness is based on at least two

reasons. First, awareness of scientific activities is not the primary task of the user but. For example, in the activity awareness scenario (Section 7.3), Diane wants to be alerted of status updates related to her citations or new papers; seldom will these be her primary activities. Because awareness of scientific activities is secondary to a user's primary task, it needs to be conveyed in a lightweight, non-intrusive way, yet be effective enough to capture the user's attention and cause some response. Notification systems fit exactly this profile. According to McCrickard and Chewar (2003), the success of a notification system hinges on accurately supporting attention allocation between tasks, while simultaneously enabling utility through access to additional information.

Second, survey results indicated that flexibility is required in configuring not only *how* awareness information should be conveyed but also *what* should be conveyed. For example, some CiteSeer users would be interested in citations to their papers, others in when new papers are available, yet some would want to know when a specific discussion thread has been posted. Notification systems provide such configurability so users get the right kind of information in the ways that they want it.

We have been exploring the use of object-based RSS feeds, as opposed to traditional news-based feeds, as notification systems to convey activity awareness in collaborative settings (Hylton et al., 2005). RSS feeds seem appropriate for conveying scientific activity awareness because of their configurable, non-intrusive, and light-weight nature. Also, many survey respondents provided strong support for RSS feeds in CiteSeer as indicated by our results.

Example. We are implementing notification systems based on RSS for supporting scientific activity awareness (Farooq et al., 2007). To get feedback from CiteSeer users, we have simulated the RSS feeds and are asking users to rate them on their usefulness and relevance. The three types of simulated feeds being evaluated are highlighted in Fig. 4: (1) *Citations*: citations to one's papers in CiteSeer; (2) *Papers with keywords*: CiteSeer papers related to one's specified keywords; and (3) *Related papers*: related papers to one's papers in CiteSeer.

7.5. Summary

Based on our survey data and current HCI/CSCW literature, we presented four implications for designing the CiteSeer collaboratory. We illustrated these design strategies with examples from our current work. Table 2 summarizes these findings.

8. Discussion

In this paper, we have developed a design agenda from a requirements survey. Such a mapping from requirements to

Table 2
Implications for designing the CiteSeer collaboratory

Design strategy	Example illustrating design
Visualize query-based social networks to identify scholarly communities of interest.	Discovering cross-community bridges from a social network.
Provide online collaborative tool support for upstream stages of scientific collaboration.	Collaborative workspace to be integrated with CiteSeer.
Support activity awareness to stay cognizant of online, asynchronous, and long-term scientific activities.	Assessing impact and influence patterns of one's research.
Use notification systems to convey scientific activity awareness peripheral to users' primary task.	Feed-based notifications to stay aware of interested papers and citations.

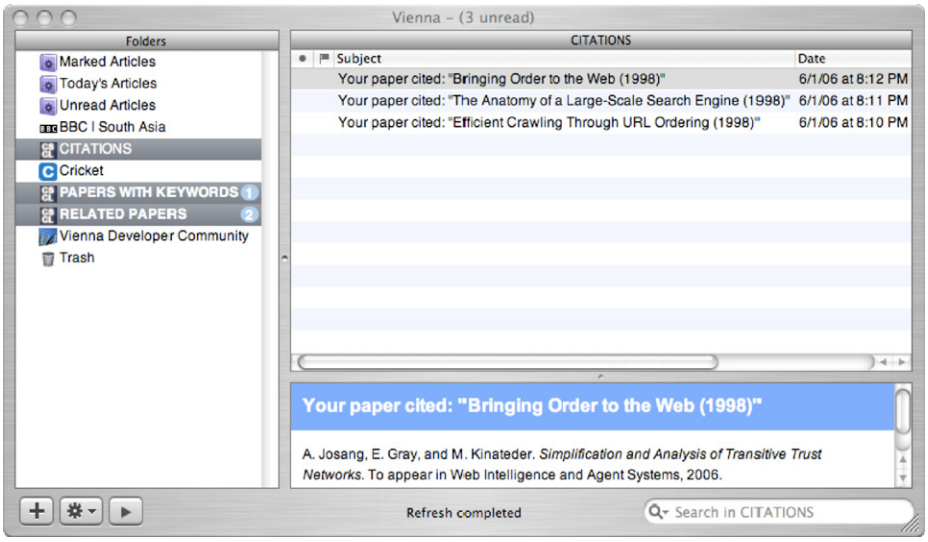

Fig. 4. RSS simulations for various publication events in CiteSeer.

design necessarily involves some inference and projection, and is inherently risky. At the same time, however, it also provides concrete and grounded guidance for design explorations.

Reflecting on our study, we briefly discuss the intellectual boundaries of our findings from our investigation into requirements for collaboration in e-science in the context of CiteSeer. Having acknowledged these limitations and alluded to future work, we draw out broader implications for our e-science vision as it applies to collaboratories around digital libraries.

8.1. Limitations of study and future work

We acknowledge that our results and implications are based on participants indicating their preferences for enhancing CiteSeer and not on their actual experiences of using the CiteSeer collaboratory. As a requirements gathering exercise targeted toward distributed users from all over the world, a survey-based study seems feasible and appropriate. In general, our follow-up studies in which CiteSeer users experienced various collaboratory prototypes corroborate the requirements presented in this paper (e.g., see Farooq et al., 2007).

It is appropriate to emphasize two caveats regarding the survey sample. First, we used an opportunity sample that may or may not be representative of the CiteSeer scientific community. The opportunity sample, based on self-selection, was however the only realistic sampling procedure available to us. There is no way to meaningfully and randomly identify a sample of CiteSeer users because the population of users is not enumerated anywhere. One becomes a user merely by accessing CiteSeer services. By posting our invitation on the main page, we in effect sampled *all* CiteSeer users who accessed the services via the main page during the survey timeframe, though we know that many did not accept our invitation. In research such as ours on Internet-based services, opportunity samples are an inescapable risk (for detailed discussion, see Coomber, 1997).

Second, in this initial survey study, we did not parse the survey data according to demographic factors. Collecting, stratifying, and analyzing data by professional status, gender, educational background, and geographical location can enrich our interpretation. For example, scientists at different stages of their careers (e.g., graduate students versus established professors) may have different needs for making new contacts and engaging in collaboration (Bruckman and Jenson, 2002). These needs would then have to be supported in a personalized way based on user profiles. We plan to take such factors into account in future CiteSeer studies.

An obvious limitation is that our investigation only considered CiteSeer. Nevertheless, we believe our findings can be applied to other *niche* digital library and search engine services. In contrast to general-purpose engines such as Google or Yahoo, niche or specialized engines, such as CiteSeer, scour cyberspace with the goal of indexing only a small subset of documents relevant to a community (Kruger et al., 2000). Kruger et al. (2000) say that the power of customized search engines derives from the fact that the underlying domain is constrained, and documents within this community have common elements. By modeling and extracting these elements, complex queries that take advantage of the sophistication of the community members within their field can be implemented.

For example, we expect our findings to generalize most directly to other niche digital libraries such as the ACM and IEEE Digital Library. Indeed, it is interesting to speculate about the apparent failure of the document-specific discussion forums in the ACM Digital Library with respect to our requirements analysis. As of December 2005, there were only 14 discussion threads and 16 total contributions to these threads. These are shockingly low numbers, given that ACM Digital Library is one of the largest collections of computer science literature. It is notable that not one of the design implications we identified is implemented in the ACM Digital Library.

As immediate future work, we are enhancing CiteSeer's infrastructure to support some of the design strategies presented in this paper (Table 2). Using examples through design envisionment scenarios, conceptual schemas, and prototype screenshots, we illustrated various threads of design implementation that need to follow this requirements gathering phase. We also identified BRIDGE (example described in Section 7.2) as a compatible environment to integrate with CiteSeer. BRIDGE already supports asynchronous, collaborative activities such as brainstorming, white-boarding, and concept mapping (some features were illustrated in Fig. 3), and provides activity awareness through notification systems in context of collaborative work (Carroll et al., 2003). By gearing the BRIDGE functionality toward scientific collaboration, we plan to iteratively prototype the CiteSeer collaboratory and formatively evaluate it with CiteSeer users.

8.2. Application of findings in the broader context of e-science usability research

Although digital libraries are often integrated with collaboratories (Finholt, 2002), they have not typically been investigated from the perspective of communities. Digital libraries are repositories for information search and retrieval, but they are also collective resources that attract people and help to form scholarly communities. Users are doing more than visiting a web site, they are building social networks, sharing knowledge, and more. They are participating in online scientific *communities of practice*.

A community of practice involves a set of socially defined ways of doing things in a specific domain (e.g., computer and information science): a set of common approaches and shared standards that create a basis for action, communication, problem solving, performance, and accountability (Wenger, 1998). A community of

practice has three fundamental elements: a *domain of knowledge*, which defines a set of issues; a *community of people* who care about this domain; and the *shared practice* that they are developing to be effective in their domain (Wenger et al., 2002). A community of practice is a medium for professional practice and development. It is also a network of individuals in various types of social and professional relationships. Newcomers gain access to the community's professional knowledge, tools, and social norms through participation in authentic activities and communication with other community members who represent a range of roles and expertise. As new members of the community gain greater expertise in the practice, their roles and positions in the community change, and they themselves begin to guide newcomers, participate in new forms of professional activities, innovate the practice, and become more central in the activity of the community.

The CiteSeer population is an implicit community of practice. CiteSeer users have the basic characteristics of a community of practice—domain of knowledge, community of people, and shared practice—but they do not have any online mechanism in CiteSeer that allows them to see, stay aware of, and interact with one another. The most significant result from our CiteSeer findings is that users want to collaborate around the intellectual resources of a digital library in ways similar to that in an online community of practice. A corollary to this result is that scientific communities around digital libraries can be better supported by tools that reinforce their identity as a community and provide an incubating environment to collaborate with others in the community. This appropriately merges e-science usability research with online communities of practice.

An exception to digital libraries not being investigated from the perspective of a community (not necessarily a community of practice) is the work by Renda and Straccia (2005). Their vision is that digital libraries can indeed be considered as collaborative meeting and common working places where users may become aware of each other, open communication channels, and exchange information and knowledge with each other or with experts. Although Renda and Straccia's contribution is purely technical in nature and does not specifically consider the social aspects of scientific communities of practice, their underlying premise that online communities need a gathering place of some sort is well-acknowledged. For example, Kim (2000) says that for distributed online communities, a gathering place can be a mailing list, a chat room, a virtual world, a blog, or some combination of these spaces. Online gathering places, just like their geographical counterparts, nourish relationships, develop a sense of community, and promote social interactions. For CiteSeer and similar infrastructures, we think there is value in letting users know they are there with others and are (or can be) part of an online community of practice. This ties back to Kim's (2000, p. 27) assertion that in order to build a successful online community, designers will need to set up gathering places that meet the needs of the target audience. CiteSeer users have conveyed these needs for building community capacity and communality through the administered survey as part of the requirements gathering process.

Given that users want to collaborate around the intellectual resources of digital libraries (as in the case of CiteSeer), it is then worthwhile to consider the nature of these intellectual resources that lead to collaboration. On the surface, these intellectual resources are essentially data (scholarly papers) and meta-data (information related to scholarly papers such as citation counts) that are part of the information search and retrieval process. For instance, CiteSeer users query, access, download, and share scholarly papers with their peers.

However, in context of e-science, the intellectual resources of scientific digital libraries are not just discrete information that can be just stored and passed from sender to receiver. These intellectual resources convey meaning to their users, foster dialectical behavior, and enable co-construction of interpretation. These intellectual resources are artifacts that foster social interaction. Such a perspective has been recently acknowledged in e-science literature. In their eDiaMoND case study of collaboration and trust in healthcare systems, Jirotka et al. (2005) suggest that requirements gathering processes for e-science would benefit from a conceptualization of "data" that goes beyond the "commodity" view of information. They suggest that an understanding of the *contextual* nature of data could inform the design of e-science systems with an emphasis on data sharing and possibly collaboration. This will be a central focus of our future research.

9. Conclusion

Since its inception, the World Wide Web has changed the ways scientists communicate, collaborate, and educate (Berners-Lee et al., 2006). With such a proliferation of the Internet, Finholt (2002) rightly points out that collaboratories represent an important convergence of computing technology with scientific practice. However, to qualitatively advance scientific practice, the design space requires novel and specific insights from more collaboratory case studies. Previous findings (Table 1) certainly inform us of general factors that hamper successful distributed collaboration (e.g., lack of common ground, collaboration readiness, etc.) but they seldom indicate specific design recommendations or strategies to counter these issues. Our implications for design, emerging from requirements gathering of the proposed CiteSeer collaboratory, extend current findings in that they can be implemented, empirically evaluated, and refined.

The empirical results from the administered survey and follow-up interviews specifically raise issues of community building and collaboration for CiteSeer users. We found that users are inclined to interact with potential collaborators based on various criteria. Making such criteria visible, such as users having similar research interests through

query-based social networks, can facilitate more meaningful collaboration.

We also reported that supporting collaborative activities that are in the early, divergent (upstream) stages of scientific discovery is a first approximation to enable collaboration currently between CiteSeer users. This can avoid issues such as trust and privacy for the time being until users become progressively motivated and confident in using collaborative tools that support more focused activities.

Finally, we reported that users perceive CiteSeer as a resource for keeping aware of the vector of activities occurring in their field and others. Activity awareness through notification systems is a promising candidate for keeping track of long-term changes to intellectual resources and shared activities around those resources.

Our findings complemented with design examples set a research trajectory for enhancing CiteSeer as a collaboratory. Some of these design examples we allude to have been implemented in other contexts. For example, online bookstores provide social networks based on what people search for (e.g., Amazon); digital libraries provide awareness of citation statistics (e.g., Google Scholar); and listservs provide notification services for recent news (e.g., ACM SIGCHI). While these features exist independently of each other, they have not been investigated in context of collaboratories as an integrated design to support distributed scientific collaboration. Also, our findings are grounded in systematic, empirical-based e-science research.

While we acknowledge that implementing and evaluating the proposed CiteSeer collaboratory will be useful (as we suggest in our future work), it is imperative to simultaneously recognize that requirements gathering is a valuable component in and of itself for investigating e-science usability research. As Ackerman (2000) puts it, "If CSCW (or HCI) merely contributes 'cool toys' to the world, it will have failed its intellectual mission". As we suggest in our related work, current findings in collaboratory design can certainly be enriched by focusing on systematic requirements gathering investigations that lead to design.

Acknowledgments

This complete paper extends an earlier, shorter version ("Supporting distributed scientific collaboration: Implications for designing the CiteSeer collaboratory") that was presented at the Hawaii International Conference on System Sciences (January 3–6, 2007; http://www.hicss.hawaii.edu/). Our work is partially supported by the US National Science Foundation (NSF) grant CRI-0454052.

References

Ackerman, M., 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. Human–Computer Interaction 15 (2/3), 181–205.

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., Weitzner, D.J., 2006. Creating a science of the Web. Science 313, 769–771.

Bruckman, A., Jenson, C., 2002. The mystery of the death of MediaMoo: seven years of evolution of an online community. In: Renninger, K.A., Shumar, W. (Eds.), Building Virtual Communities: Learning and Change in Cyberspace. Cambridge University Press, pp. 21–33.

Carroll, J.M., Neale, D.C., Isenhour, P.L., Rosson, M.B., McCrickard, S.D., 2003. Notification and awareness: synchronizing task-oriented collaborative activity. International Journal of Human–Computer Studies 58, 605–632.

Carroll, J.M., Rosson, M.B., Convertino, G., Ganoe, C.H., 2006a. Awareness and teamwork in computer-supported collaboration. Interacting with Computers 18 (1), 21–46.

Carroll, J.M., Rosson, M.B., Farooq, U., Convertino, G., Merkel, C.B., Schafer, W.A., Ganoe, C.H., Xiao, L., 2006b. Beyond being aware. Human–Computer Interaction Consortium (Fraser, Colorado, February 1–5, 2006).

Clark, H.H., Brennan, S.E., 1991. Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (Eds.), Perspectives on Socially-shared Cognition. American Psychological Association, Washington, DC, pp. 127–149.

Coomber, R., 1997. Using the Internet for survey research. Sociological Research Online 2 (2).

Cummings, J., Kiesler, S., 2005. Collaborative research across disciplinary and institutional boundaries. Social Studies of Science 35, 703–722.

David, P.A., 2006. Toward a cyberinfrastructure for enhanced scientific collaboration: providing its "soft" foundations may be the hardest part. In: Kahin, B., Foray, D. (Eds.), Advancing Knowledge and the Knowledge Economy. MIT Press, Cambridge, pp. 431–453.

De Roure, D., Jennings, N., Shadbolt, N., 2001. Research agenda for the semantic grid: a future e-science infrastructure. Technical Report UKeS-2002-02, National e-Science Centre, Available at: 〈http://www.nesc.ac.uk/technical_papers/DavidDeRoure.etal.SemanticGrid.pdf〉.

Dourish, P., 2001. Where the Action Is: The Foundations of Embodied Interaction. MIT Press, London.

Dourish, P., Bellotti, V., 1992. Awareness and coordination in shared workspaces. In: Proceedings of the Conference on Computer Supported Cooperative Work, Toronto, Canada, October 31–November 4, 1992. ACM Press, New York, NY, pp. 107–113.

Erickson, T., Smith, D.N., Kellogg, W.A., Laff, M.R., Richards, J.T., Bradner, E., 1999. Socially translucent systems: social proxies, persistent conversation, and the design of Babble. In: Proceedings of the Conference on Human Factors in Computing Systems, Pittsburg, PA, May 15–20, 1999. ACM Press, New York, NY, pp. 72–79.

Farooq, U., 2006. Eureka! Past, present, and future of creativity research in HCI. ACM Crossroads 12 (3), 33–35 (Spring).

Farooq, U., Carroll, J.M., Ganoe, C.H., 2005. Supporting creativity in distributed scientific communities. In: Proceedings of the International GROUP Conference on Supporting Group Work, Sanibel Island, Florida, November 6–9, 2005. ACM Press, New York, pp. 217–226.

Farooq, U., Ganoe, C.H., Carroll, J.M., Councill, I.G., Giles, C.L., 2007. Design and evaluation of awareness mechanisms in CiteSeer. Information Processing and Management, in press, doi:10.1016/j.ipm.2007.05.009.

Finholt, T.A., 2002. Collaboratories. In: Cronin, B. (Ed.), Annual Review of Information Science and Technology, vol. 36. ASIST, Washington, DC, pp. 73–107.

Finholt, T.A., Olson, G.M., 1997. From laboratories to collaboratories: a new organizational form for scientific collaboration. Psychological Science 8 (1), 28–35.

Foster, I., Kesselman, C. (Eds.), 1998. The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco, CA.

Ganoe, C.H., Somervell, J.P., Neale, D.C., Isenhour, P.L., Carroll, J.M., Rosson, M.B., McCrickard, D.S., 2003. Classroom BRIDGE: using collaborative public and desktop timelines to support activity awareness. In: Proceedings of UIST, Vancouver, Canada, Nov 2–5, 2003, ACM Press, New York, pp. 21–30.

Giles, C.L., Bollacker, K., Lawrence, S., 1998. CiteSeer: an automatic citation indexing system. In: Proceedings of the Conference on Digital Libraries, Pittsburg, PA, June 23–26, 1998. ACM Press, New York, NY, pp. 89–98.

Goodrum, A.A., McCain, K.W., Lawrence, S., Giles, C.L., 2001. Scholarly publishing in the Internet age: a citation analysis of computer science literature. Information Processing and Management 37, 661–675.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.

Gutwin, C., Greenberg, S., 1996. Workspace awareness for groupware. In: Proceedings of the Conference on Human Factors in Computing Systems, Vancouver, Canada, April 13–18, 1996. ACM Press, New York, NY, pp. 208–209.

Herlocker, J.L., Konstan, J.A., Riedl, J., 2000. Explaining collaborative filtering recommendations. In: Proceedings of the Conference on Computer Supported Cooperative Work, Philadelphia, PA, December 2–6, 2000. ACM Press, New York, NY, pp. 241–250.

Hinde, S., Wilcock, S., 2002. The GRID as a platform for communication, collaboration, and e-science. Technical report HPL-2002-125, Technical Publications Department, HP Labs Library, San Francisco, CA.

Hollan, J., Stornetta, S., 1992. Beyond being there. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Monterey, CA, May 3–7, 1992. ACM Press, New York, NY, pp. 119–125.

Hylton, K., Rosson, M.B., Carroll, J.M., Ganoe, C.H., 2005. When news is more than what makes headlines. ACM Crossroads 12(2). Available online: 〈http://www.acm.org/crossroads/xrds12-2/rss.html〉.

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C., Voss, A., 2005. Collaboration and trust in healthcare innovation: the eDiaMoND case study. Computer Supported Cooperative Work 14, 369–398.

Kavanaugh, A., Reese, D.D., Carroll, J.M., Rosson, M.B., 2005. Weak ties in networked communities. The Information Society 21 (2), 119–131.

Kelly, S.U., Sung, C., Farnham, S., 2002. Designing for improved social responsibility, user participation and content in on-line communities. In: Proceedings of Conference Human Factors and Computing Systems, Minneapolis, MN, April 20–25, 2002. ACM Press, New York, NY, pp. 391–398.

Kim, A.J., 2000. Community building: Secret Strategies for Successful Online Communities on the Web. Peachpit Press, Berkeley, CA.

Kouzes, R.T., Myers, J.D., Wulf, W.A., 1996. Collaboratories: doing science on the Internet. IEEE Computer 29 (8), 40–46.

Kruger, A., Giles, C.L., Coetzee, F., Glover, E., Flake, G.W., Lawrence, S., Omlin, C.W., 2000. Deadliner: building a new niche search engine. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, Virginia, November 6–11, 2000. ACM Press, New York, pp. 272–281.

Levine, J.M., Moreland, R.L., 2004. Collaboration: the social context of theory development. Personality and Social Psychology Review 8 (2), 164–172.

McCrickard, D.S., Chewar, C.M., 2003. Attuning notification design to user goals and attention costs. Communications of the ACM 46 (3), 67–72.

McCrickard, D.S., Chewar, C.M., Somervell, J.P., Ndiwalana, A., 2003. A model for notifications systems evaluation—assessing user goals for multitasking activity. ACM TOCHI 10 (4), 312–338.

McDonald, D.W., 2001. Evaluating expertise recommendations. In: Proceedings of the International GROUP Conference on Supporting Group Work, Boulder, CO, September 30–October 3, 2001. ACM Press, New York, pp. 214–223.

Olson, G.M., Olson, J.S., 2001. Distance matters. Human–Computer Interaction 15, 139–179.

Olson, G.M., Finholt, T.A., Teasley, S.D., 2000. Behavioral aspects of collaboratories. In: Koslow, S.H., Huerta, M.F. (Eds.), Electronic Collaboration in Science. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–14.

Olson, J.S., Hofer, E.C., Bos, N., Zimmerman, A., Olson, G.M., Cooney, D., Faniel, I., 2008. A theory of remote scientific collaboration (TORSC). In: Olson, G.M., Zimmerman, A., Bos, N. (Eds.), Scientific Collaboration on the Internet. MIT Press, Cambridge, MA.

Renda, M.E., Straccia, U., 2005. A personalized collaborative digital library environment: a model and an application. Information Processing and Management 41, 5–21.

Sargent, M.A., 2004. Report of a mission to the UK to examine the UK e-science programme in the context of Australian e-research initiatives. Technical report, The Australian Academy of Technological Sciences and Engineering. Available at: 〈www.atse.org.au/uploads/escience.pdf〉.

Science of Collaboratories, 〈http://www.scienceofcollaboratories.org〉 (accessed 16.03.2006).

Sonnenwald, D.H., Whitton, M.C., Maglaughlin, K.L., 2003. Evaluating a scientific collaboratory: results of a controlled experiment. ACM TOCHI 10 (2), 150–176.

Star, S.L., Ruhleder, K., 1994. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In: Proceedings of the Conference on Computer Supported Cooperative Work, Chapel Hill, NC. ACM Press, New York, NY, pp. 253–264.

Star, S.L., Ruhleder, K., 1996. Steps towards an ecology of infrastructure: design and access for large information spaces. Information Systems Research 7 (1), 111–134.

Sutton, R.I., Hargadon, A., 1996. Brainstorming groups in context: effectiveness in a product design firm. Administrative Science Quarterly 41, 685–718.

Terveen, L., McDonald, D.W., 2005. Social matching: a framework and research agenda. ACM TOCHI 12 (3), 401–434.

Wellman, B. (Ed.), 1999. Networks in the Global Village: Life in Contemporary Communities. Westview Press, Boulder, CO.

Wenger, E., 1998. Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press, New York.

Wenger, E., McDermott, R., Snyder, W., 2002. Cultivating Communities of Practice: A Guide to Managing Knowledge. Harvard Business School Press, Boston, MA.

Wulf, W.A., 1993. The collaboratory opportunity. Science 261, 854–855.