ECSCW 2003
Proceedings of the Computer Supported Scientific Collaboration Workshop, Eighth European Conference on Computer Supported Cooperative Work, Helsinki, Finland, 14 September 2003

Helena Karasti, Karen Baker, Geoffrey C. Bowker (Editors)

University of Oulu
Department of Information
Processing Science
Research Papers Series: A34

ISBN: 951-42-7121-1
ISSN: 0786-8413

Preface

The Eighth European Computer Supported Cooperative Work conference (ECSCW 2003), a biannual forum that brings together the social and technical aspects for supporting collaborations, provides a venue this year to gather researchers interested in the study of scientific collaborations and their technology support.

The response to the call for papers for the Computer Supported Scientific Collaboration workshop (CSSC) is evidence of the CSCW community members shared interest in the elements of collaboration particular to scientific communities and in the challenges they present for designing computer-based support systems.

Like the Three Smiths of Nylund's statue in Helsinki, we three organizers came together to work jointly at crafting an understanding of a scientific network, the Long-Term Ecological Research program. Having hammered out a workshop agenda, we welcome the CSSC participants. Through the position papers collected here and with our diverse case studies, from ecological teams and human genes to digital streams, we add shape to the characterization of CSSC.

Kiiminki, September 7 2003

Helena Karasti, Karen S. Baker and Geoffrey C. Bowker

Contents

Call for Position Papers: Computer Supported Scientific Collaboration Workshop, ECSCW'03, September 14-18 2003, Helsinki Finland

Karen S. Baker, Geoffrey C. Bowker and Helena Karasti
http://ecscw2003.oulu.fi/

Workshop description

The predominant interest within CSCW has been on forms of work other than scientific collaboration. Some research communities, such as Science and Technology Studies, have a long tradition of studying scientific work and recently examples can be found also in CSCW conferences and literature. Scientific collaboration differs in several ways from the business communities which CSCW has centered on. Today's scientific work can be characterized as orchestrating responses to an explosion of data leading to 'phenomenal amounts of data'. This in turn increases and intensifies multi- and interdisciplinary collaboration necessitating continuing negotiations between participants, organizations, and disciplines on issues of coordination, attribution, identity, standards, protocols, and data sharing. Discipline specific as well as organizational and sociotechnical informatics are emerging to address such issues. Since scientific work is heterogeneous, different fields and interdisciplinary collaborations face varied challenges and therefore also pose diverse challenges for CSCW. For example, and to make a comparison, both molecular biology and long term ecology deal with vast amounts of data. The field of molecular biology is in period of remarkably rapid change, as the genome sequencing projects and new experimental technologies have generated an explosion of data (O'Day et al. 2001). Long-term ecology, in turn, deals with biodiversity issues and faces the challenge of data diversity (Bowker 2000). Both must address

problematics of the enduring long-term information management and infrastructure (NSF 2003; Baker et al. 2002). Understanding how to facilitate communication and collaboration through software, databases, and infrastructure is essential for both areas of science given the variety of different, often unarticulated issues at stake.

Goals and objectives

The aim of the workshop is to bring together, for the first time in the CSCW venue, researchers who share an interest in the study of scientific collaborations and their technology support:

- to map current research initiatives;
- to create relations among diverse researchers;
- to identify areas of mutual interest and to scope the challenges involved in scientific collaborations for the development of technological support; and
- to explore plans for a follow-on activity such as a joint publication forum in a journal special issue.

Activities and discussions

The workshop will be organized around a number of themes. These are by no means meant to be exclusive topics but will be extended to take account of emergent themes based on the contributions of the participants.

- Is scientific collaboration different from other types of collaborative work? How does it differ?
- What are the varieties of scientific collaborations studied? Do they however have some commonalities (in comparison with other areas of work)?
- What are characteristic of the relations between scientific collaboration as 'use practice' and 'technology design'?
- What are the challenges scientific collaborations pose for CSCW?

Participation

We are looking for participants with diverse backgrounds and interests in the area of computer support for scientific collaboration. People interested in participating are requested to submit position papers not exceeding four pages to Helena Karasti (helena.karasti@oulu.fi) by June 16th 2003. Since there is a limit to how many can interact in a workshop-style event, we will limit the number of participants to approximately fifteen. The workshop organizers will review the position papers and select the most promising for the workshop.

Organizers

The organizers started to work together in 2002 on a NSF funded BioDiversity and EcoInformatics (BDEI) project "Designing an Infrastructure for Heterogeneity in Ecosystem Data, Collaborators and Organizations" (http://pal.lternet.edu/projects/02dgo/). Having encountered a multitude of important issues that benefit from an interdisciplinary team approach, the collaboration continues.

Karen S. Baker (http://www.icess.ucsb.edu/~karen/) is a member of Scripps Institution of Oceanography at University of California, San Diego and the information manager for the Palmer Long-Term Ecological Research (LTER) site studying an Antarctic marine ecosystem. Both the Palmer LTER (a team of researchers distributed at institutions across the United States) as well as the LTER network (a federation of sites representing 24 different ecosystems) collaborate scientifically supported by cooperative infrastructure and information management (Baker et al. 2000). Karen is interested in the design of infrastructure to promote the flow of information between people and through time.

Geoffrey C. Bowker (http://weber.ucsd.edu/~gbowker) chairs the Department of Communication at the University of California, San Diego. His studies have focused on the development of information infrastructures and their relationship with knowledge (Science on the Run: Industrial Geophysics and Information Management, 1994) and classification systems (Sorting Things Out: Classification and its Consequences, 1999), broadening to consideration of sociotechnical and organizational elements of infrastructures. His interests include distributed scientific practices and collaborative science.

Helena Karasti (http://www.tol.oulu.fi/~helena/) is currently acting professor at University of Oulu, Finland. Last year she visited UCSD and conducted extensive fieldwork within the Long Term Ecological Research (LTER) network. Helena is interested in technologically mediated work, relations of digital and material mediation, everyday practices and expertises, and the relations of use and design. Her recent doctoral dissertation (2001) 'Increasing sensitivity towards everyday work practice in system design' explores the integrations of ethnographic studies of work practice and participatory design (http://herkules.oulu.fi/isbn9514259556/).

Important dates

Submission of position papers: June 16th 2003 (to Helena.Karasti@oulu.fi)
Notification of acceptance: July 7th 2003
Workshop: September 14th 2003

References

Baker, K., B. Benson et al. (2000). Evolution of a Multisite Network Information System: the LTER Information Management Paradigm. BioScience, 50(11), pp. 963-968.

Baker, K., G. C. Bowker & H. Karasti (2002): Designing an Infrastructure for Heterogeneity in Ecosystem Data, Collaborators and Organizations, DG.O 2002 National Conference for Digital Government Research, May 19-22, 2002, Los Angeles, CA, USA.

Bowker, G. C. (2000): Biodiversity Datadiversity. Social Studies of Science (30):643-683.

Bowker, G. C. & S. L. Star (1999): Sorting Things Out: Classification and Its Consequences. MIT Press, London.

Karasti, H., K. S. Baker & G. C. Bowker (2003): Ecological Storytelling and Collaborative Scientific Activities. SIGGROUP Bulletin, in press.

NSF (2003): Revolutionizing Science and Engineering through Cyberinfrastructure. http://www.communitytechnology.org/nsf_ci_report/

O'Day, V., A. Adler et al. (2001): When worlds collide: Molecular biology as interdisciplinary collaboration. In Proceedings of ECSCW'01. pp. 419-418.

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

Contact

Helena Karasti
Department of Information Processing Science
University of Oulu
P.O.Box 3000
FIN-90014 Oulu University
FINLAND
Email: Helena.Karasti@oulu.fi

CSSC Workshop Agenda

INTRODUCTIONS
Introductions to the day (Baker & Karasti)
Participant introductions (name, organization, background, science collaborations you are familiar with)

TALK: Introduction to CSSC (Bowker)
GROUP ACTIVITY 1 – CSSC Themes: Each participant identifying some common threads in position papers
COMMENT: Participants' impressions

BREAK

TALKS A: Scientific practices (Fry, Meier, Talja)
GROUP ACTIVITY 2 – SC characteristics: Each participant providing some keywords describing scientific collaborations
COMMENT: Participants' impressions

LUNCH

TALKS B: Bridging Scientific Practices and Design (Cash, Koschmann)
GROUP ACTIVITY 3 – CSSC design issues: Each participant providing some critical issues in bridging technology development and work practices for scientific collaborations
COMMENT: Participants' impressions

BREAK

GROUP ACTIVITY – Integration of Group Activities 1-3

FOLLOW-ON PLANS

WRAP-UP AND ADJOURN

Managing Information for Mass Fatality Identification: Gene Code Forensics and the World Trade Center Disaster

Debra Cash and Howard Cash
Gene Codes Forensics, Ann Arbor, Michigan, USA

In late September, 2001, Gene Codes Forensics, a bioinformatics company based in Ann Arbor, Michigan, took on the challenge of creating a software application that could track the remains, personal effects, kinship data and DNA of the victims of the World Trade Center attack in order to identify the victims. This has been the most ambitious forensics identification effort in history.

Gene Codes established five goals:

- Identify individual remains
- Reunify partial remains so that they can be returned to families
- Collect and warehouse meta-data for administrative review of reference samples (antemortem victim materials, such as toothbrushes, razors etc.)
- Track samples among collaborating laboratories
- Create an information management system to report metrics and make problem resolution proposals to supervisors at the New York City Office of Chief Medical Examiner (OCME)

The programs available to the OCME at the time of the disaster (primarily CoDIS, the Federal Bureau of Investigation's Combined DNA Index System) had been designed around unique identifiers, such as fingerprints, and "clinical draws" in which

DNA information from bone marrow, blood samples or cheek swabs was unambiguous. CoDIS had no way to group and collapse data from dozens of fragmented remains (in one unfortunate case, a victim was fragmented into more than 200 pieces) or perform "all against all" matches. Nor could it accommodate data from degraded DNA samples (as the world knows, "Ground Zero" was on fire for over three months) and incomplete genetic profiles. CoDIS was also

inadequate to deal with DNA profiles generated by commingled remains, crushed together under the weight of the towers' collapse.

Gene Codes is best known for its market-leading DNA sequencing product, Sequencher. (See www.genecodes.com for more details.) That product was designed with a profound commitment to user-centered design and an understanding the work practices of bench scientists in a number disciplines. The OCME, as well as the US Army and other government agencies, were already customers of Gene Codes on September 11, 2001, and turned to the company for support within days of the disaster.

Gene Codes Forensics embarked on a complete — and difficult — ethnographically-informed workplace mapping and inventory assessment, tracking how information flowed among bench scientists, work groups, computer systems (such as installed laboratory management systems) and participating labs, how information was tracked and transformed at each state, and how it was reconciled at the point of identification and confirmed before identifications were released to medical-legal investigators and the victims' families.

There was special concern with understanding how information gathered — and mistakes made — at one point in the process could have deleterious efforts on the ease or accuracy of identification later in the process. (As a simple example, in the first two weeks after the WTC disaster, the New York State Police or other subcontracted agencies received over 12,000 individual items, such as toothbrushes, razors and hairbrushes. This astonishing flow of materials overwhelmed the processes that had been put in place to manage it on an emergency basis. Later, an "administrative review" process — primarily a paper-based research process — would have to be put in place to ascertain that "John Doe's" toothbrush actually *belonged* to John Doe and had not been donated by a the family of another missing person.)

M-FISys (pronounced "Emphasis"), the product developed for the WTC identification effort, had to address — and resolve — a number of issues of special importance to CSCW researchers, especially those working in the biological sciences. These include:

- M-FISys had to be a "tool for skilled work." The software application itself does not "make" identifications: only certified forensic scientists with a certain level of expertise may make identifications and approve the issuing of death certificates.
- M-FISys had to display and integrate information coming from a number of sources: the medical examiner's office, the state police forensic lab which was handling personal effects and family references, plus a number of high-throughput commercial laboratories. On September 11, 2001 these organizations not only had incompatible applications, they had incompatible networks and in most cases, incompatible and often confounding nomenclature schemes that only became more baroque over time.

- M-FISys had to allow the forensic scientists to interrogate the raw data behind genetic profiles generated by the participating labs and reported in the system database
- MFISys had to build in algorithms to encode legally-designated kinship likelihood ratios, an exemplar of instantiating or "pointing to" extrinsically defined scientific standards in any scientific practice
- M-FISys was developed iteratively in tandem with rapidly changing work conditions and scientific practice. Although the particular group of end-users (staff criminalists) in the Medical Examiner's office remained relatively stable, their activities changed, in some ways dramatically, over time, in some cases because the M-FISys software offered them new capabilities. Using Extreme Programming (XP) methodologies, where programmers work in pairs and unit tests and acceptance tests are written before new functionality can be added, the Gene Codes Forensics engineers deployed new releases of M-FISys on a weekly basis, starting in December 2001. (There have been more than 70 releases of the product as of June 2003.) New functionality was added after observation and direct negotiation with users about their most pressing priorities for any given period of time. In addition, the use of Single Nucleotide Polymorphism (SNP) from nuclear DNA, while accepted in genetic research, had never been used in forensic identification efforts. It is being added to the arsenal of identification modalities in the hope that this will help identify persons whose remains are otherwise unidentifiable. (As I write this, I believe the use of SNPs has not yet been certified for WTC identification by the national agencies involved; this information, however, has already been referred to in media reports).

The development of M-FISys is a compelling case study of a number of key CSCW concerns:

- How do we design and build computer systems under conditions of incomplete information and changing workplace conditions?
- How can systems be developed quickly in settings where *there cannot be a single mistake?*
- What are the key differences between building a system for an identified group of beta users or "early adopters" (as are most of the cases of systems built under academic research efforts) and commercial development (such as Gene Codes' Sequencher product)?
- What, if anything, are the salient differences between designing for users who will employ a CSCW system under day to day conditions (here, making identifications) and those who will be rolling up cumulative or overview data in report format (such as the lab managers reporting to the mayor and other officials)?
- What, if anything, are the design and technical concerns associated with systems whose *deliverables* (in this case, confirmed identifications of

victims) will ultimately be delivered to people other than the end-users (in this case, the families of victims, rather than the forensic scientists themselves)? Does this have any implications for understanding large-scale systems such as applications that will have broad public policy implications in the areas of environmental sciences, epidemiology, or bioengineering?

Papers written or submitted on Gene Codes Forensics and the WTC effort

[Please note that papers on this topic were embargoed until late in 2002 and others will be forthcoming.]

Brenner, C.H. and Weir, B.S., "Issues and Strategies in the DNA Identification of World Trade Center Victims" in Theoretical Population Biology, 63 (2003) 173-178.

Cash, D., et al "Homicide x2,793: Transforming Forensics Practice into Mass Fatality Identification", under review for 2003 conference in the U.S.

Cash, H.D., Hoyle, J.W. and Sutton, A.J., "Development under Extreme Conditions: Forensic Bioinformatics in the Wake of the World Trade Center Disaster," Pacific Symposium on BioComputing, January, 2003.

Hennessey, Mike "World Trade Center DNA Identifications: The Administrative Review Process," Promega, 2002.

The Cognitive and Social Shaping of Scientific Collaboration

Jenny Fry
NERDI (Networked research and digital information), NIWI-KNAW, Amsterdam, The Netherlands

> Communication is central to the academic enterprise… [it] is the force that binds together the sociological and epistemological, giving shape and substance to the links between knowledge forms and knowledge communities. (Becher, 1989, p. 77)

The past two decades have been host to an explosion in information communication technologies (ICTs). This has created a cornucopia of digital networks and resources connected on a global scale. Scholars are no longer limited to the annual meetings of scholarly associations and societies to communicate informally with their national and international peers. They have the opportunity to stay in touch with their fields through a plenitude of email networks. Availability of channels for the formal communication of scholarly work has expanded far beyond the local collections of academic libraries. These developments have been accompanied by a great deal of speculation about the impact of digital communication media, such as the Web, on the work of scholars and the production of knowledge. We are told that with the arrival of the Internet there has been an 'information revolution' that will potentially alter scholarly communication in radical ways. However, there is a need to develop a grounded understanding of how scientists are actually using the Internet in their scholarly work.

My research has been concerned with the mutually constitutive relationship between scholarly research cultures and the use of ICTs. This paper will draw on case studies within high-energy physics and corpus-based linguistics for illustrative examples. High-energy physicists are concerned with discovering matter from which the universe is made. They do this by using large-scale apparatus, such as particle accelerators. The focus of corpus-based linguistics is the development and analysis of large corpora of examples of language in use. There has been an

increasing use of computational and statistical techniques in the development and use of corpora. The case studies were based on a series of in-depth interviews with academic researchers from universities across England. Findings from the case studies demonstrate that technology does not have an autonomous effect, but is appropriated by research communities based on their specific cultural characteristics.

Current understanding in research into scholarly communication indicates that a range of social conditions will influence the uptake and use of computer-mediated communication technologies within scholarly communities. For example, Orlikowski and Gash (in Kling and Lamb, 1996, p.48) found that "people's fine-grained work incentives influence whether they see technologies as relevant, and the ways in which they appropriate the technologies". Kling, Spector and McKim (2002) illustrate how the cultural context of disciplines can lead to the rejection of digital resources. They observed that attempts made by the National Institutes of Health (NIH) in North America to implement a digital pre-print server model of publishing (arXiv.org) in bio-medical science, which is popular within the disciplines of physics, mathematics and chaos theory, were resisted by lead scientists in the field. They believe that resistance to the model can be attributed to disparity between the pre-existing model of publishing in bio-medical science and the culture of publishing inherent in the digital pre-print server model. Olson and Olson (2000) found that existing work practices also influence the use of CMC (computer-mediated communication) technologies for collaborative work in science due to mechanisms such as reward systems. Focus amongst these studies has tended to be on the social context of scholarly practices, whereas STS (science and technology studies) scholars have long recognized the dual role of both cognitive and social considerations in shaping communication, collaboration and knowledge production (Law, 1973: Mullins, 1972: Mulkay and Edge, 1976).

Comparison of influential cultural factors across scientific fields is problematic due to the multi-faceted nature of scientific cultures. Differences in patterns of communication for collaboration have been accounted for from a variety of perspectives. For example, Kling, Spector and McKim (2002) used publishing models as a frame of reference, Olson and Olson (2000) were interested in the affects of geographic distance, and Kraut, Galegher and Egido (1988) examined the influence of interpersonal factors. All of these arguments are valid, but constrained in that they each only account for a limited number of facets of a research community's complex cultural identity. Additionally, research in the information science tradition tends to focus on comparing the final products of formal communication, such as journal articles, with behaviours online, rather than the process of knowledge production at the level of informal communication, such as workshops.

Studies such as Kraut, Galegher and Egido (1988) have made a significant contribution to CSCW because they highlight the need to consider both cognitive

(the task at hand) and social (interpersonal relations) factors in understanding collaborative processes. However, their research was predominantly focused on industrial groups. Applying an economic geography perspective to scientific collaboration Frenken and vanOort (2003) have shown that there are important differences between industrial knowledge production and scientific knowledge production. The main differences that they identified are the nature of the knowledge being produced and incentive structures. They argue that levels of tacit knowledge are lower in science and that there is a greater emphasis placed on dissemination of results. Their findings demonstrate the need to develop understandings of field differences in computer-supported scientific collaboration based on investigation of cognitive and social structures.

Whitley (2000) argues that the cognitive and social organisation of science can be conceptualised along the axes of 'task uncertainty' and 'mutual dependency'. Both of these concepts integrate cognitive and social considerations, for example Whitley stratifies 'task-uncertainty' into technical and strategic factors, and 'mutual dependence' into functional and strategic factors. 'Task uncertainty' concerns the unpredictability of task outcomes, which Whitley links to the scholarly recognition and reward system. He argues that because the sciences are committed at an institutional level to produce novel results, research activities are "uncertain compared to other work activities" in that "outcomes are not repetitious and highly predictable". 'Mutual dependence' relates to the extent to which a field is dependent upon knowledge produced in other fields in order to make a significant contribution to science and the degree of 'mutual dependence' between scientists. For example, the extent to which scientists' are dependent upon particular groups of colleagues to make competent contributions to collective intellectual goals and acquire prestigious reputations that lead to material rewards. It also accounts for the extent to which a field adopts evaluation criteria and standards from other fields for the assessment of work produced outside its intellectual boundary, rather than developing its own criteria. Whitley uses 20th Century chemistry as an example of a field that has high levels of 'mutual dependence', but low levels of 'task uncertainty', while he uses sociology as an example of low levels of 'mutual dependence', but high levels of 'task uncertainty'.

The advantage of Whitley's taxonomy over the analyses reviewed in this paper is its thoroughness in explaining the multidimensionality of scientific activity. This makes it an effective tool for exploring differences in patterns of behaviour across scientific communities. More particularly, it can be used as an explanatory framework to study the role of a field's intellectual and social organisation in the formation of collaborative work practices and the use of CMC technologies to support those practices. This argument can be illustrated with some examples from my case study data.

Intellectually, corpus-based linguistics is located at the intersection where a number of fields overlap, e.g. theoretical linguistics, computational linguistics and

natural language processing, this is a contributing factor to its fluid and diffusely bounded social organisation. Corpus-based linguists deal with a wide range of spoken languages and need to build large complex technical systems across a range of national research and technological infrastructures. This tends to result in uncertain task outcomes and a lack of standardised procedures. These characteristics match Whitley's (2000) description of a domain that has high levels of 'task uncertainty'. He argues that the implications of high-levels of 'task uncertainty' for the organisation and control of research are an increased reliance upon direct and personal control of how research is conducted, local variations in work goals and processes, and greater emphasis upon informal communication and coordination processes.

Co-ordination was a particular challenge for the widely distributed European and international collaborations concerned with building multi-lingual parallel corpora. This may account for the lack of success reported in a number of large-scale collaborative projects involving geographically distributed partners. According to some respondents the projects had failed in the sense that they had not met their objective to produce a technically functioning corpus of analytic quality within the funding period. This, however, could also be an indication that funding periods are insufficient. The corpus-based linguists reported that informal face-to-face communication was essential for coordinating European and international research projects, and that due to the wide-distribution of participants such communication was very limited. Although the corpus-based linguists recognized the importance of technical standards they had not succeeded in developing a unified system of technical and social standards and protocols for CMC across the domain. Use of the Web and other CMC technologies appears to be determined at the level of individual research groups, rather than on a community-wide basis. I argue that the lack of success across the corpus-based linguistics community in developing community-wide social and technical standards and protocols for computer-mediated collaborative work can be attributed to high levels of 'technical task uncertainty'.

Patterns of work organization and practices observed within the corpus-based linguistics case study contrast to those reported by the high-energy physicists. Coordination on a large geographic scale was also a central concern within the high-energy physics community. However, unlike the corpus-based linguists the high-energy physicists had been very successful in developing community-wide social and technical standards and protocols for the effective use of CMC to support collaborative work. An example of which is the use of the Web to transmit the live running status of experiments twenty-four hours a day seven days a week. High-energy physics fits Whitley's description of a structure that results from a field with low levels of 'task uncertainty', in that "work techniques are well understood and produce reliable results in various scientific fields", but high levels of 'mutual dependency. Whereas, Corpus-based linguistics has high levels of 'task uncertainty', with a lower degree of 'mutual dependency'. I conclude that these

differences in the cultural identity of each case study can be used to explain differences in the uptake and use of ICTs for collaborative work.

References

Becher, T. (1989). Academic tribes and territories: Intellectual enquiry and the culture of disciplines. Buckingham: SRHE & Open University Press.

Frenken, K. and van Oort, F. (2003). The geography of research collaboration in US aerospace engineering and US biotechnology & applied microbiology. Third European meeting on applied evolutionary economics (EMAEE). Augsburg, Germany, 10-12 April 2003.

Kling, R. and Lamb, R. (1996). Analyzing visions of electronic publishing and digital libraries. In: Newby, G.B. and Peek, R.M. (Eds.) Scholarly Publishing: The Electronic Frontier. Cambridge MA: The MIT Press.

Kling, R., Spector, L. and Mckim, G. (2002). Locally controlled scholarly publishing via the Internet: The Guild Model. The Journal of Electronic Publishing. **8**(1). Available at: http://www.press.umich.edu/jep/08-01/kling.html.

Kraut, R.E., Galegher, J., and Egido, C. (1988). Relationships and tasks in scientific research collaboration. Human-Computer Interaction **3**, pp. 31-58.

Law, J. (1973). The development of specialities in science: the case of X-ray protein crystallography. Science Studies **3**, pp. 275-303.

Mulkay, M.J. and Edge, D.O. (1976). Cognitive, technical and social factors in the growth of radio astronomy. In: Lemaine, G., MacLeod, R., Mulkay, M. and Weingart, P. (eds) (1976). Perspectives on the emergence of scientific disciplines. Chicago, Illinois: Aldine publishing company, pp.153-186.

Mullins, N.C. (1972). The development of a scientific specialty: The Phage Group and origins of molecular biology. Minerva **10**(1), pp. 51-82.

Olson, G.M., and Olson, J.S. (2000). Distance matters. Human-Computer Interaction **15**, pp. 139-178.

Walsh, J.P. and Bayma, T. (1996). Computer networks and scientific work. Social Studies of Science **26**, pp. 661-703.

Whitley, R. (2000). (2nd ed.) The intellectual and social organization of the sciences. Oxford: Clarendon Press.

The Data Collective: A New Model for the Ownership and Use of Scientific Data

Timothy Koschmann
Southern Illinois University, School of Medicine, USA

Traditionally, the ownership of scientific data has resided within an isolated laboratory. The laboratory itself could range from the working space of a single researcher to quite elaborate institutional structures involving many researchers, post-docs, and graduate students. Regardless, the traditional model has been one in which access to the data upon which scientific publication is based is locally controlled and tightly regulated. In most cases, the only public access to the data upon which research is based is provided in the form of published summaries and distillations. Pressures to change this model have come from a variety of quarters. Federal funding agencies have an interest in seeing data acquired at large public expense be put to maximal use. Also, various scandals involving the falsification of data in isolated laboratories has suggested the need for the development of more open data verification mechanisms. Finally, but perhaps most importantly, the notion of proprietary control of scientific data seems to clash with the view of science as an open and collaborative enterprise. As a result, a number of alternatives to the traditional model of lab-centered data ownership have emerged over time.

In certain disciplines such as linguistics there is a tradition of creating shared data corpora. Examples would include the London-Lund data corpus in communication studies, the CHILDES database in developmental psychology (MacWhinney, 1995), and the TIMSS video data in instructional science. All have international scope, both in terms of content and in use. The model here is one of collaborative effort to create a shared repository of data which then becomes a resource for multiple researchers and research groups. This is an efficient model for producing research and one that maximizes the use of the data. It is also a model that facilitates critical discussion within a field of inquiry, since the data constitutes

an object for shared attention and discussion. But the model has certain built in limitations. Shared data corpora tend to be designed for use within a particular discipline and, once constructed, tend to be largely static in structure.

The National Science Foundation in the U.S. has been instrumental in promoting another model of data sharing known as the *collaboratory*. This is a more general model for sharing scarce resources among multiple research groups and facilitating coordination among these research groups. The resources may be data, apparatus, or facilities such as observatories and small-particle accelerators. One example would be NCSA, which was initially created to provide shared access to a number of super computers. Research groups making use of a collaborator need not be co-located and researchers from different disciplines may contribute to the work of the collaboratory. The existence of a collaboratory does not necessarily require shared ownership of scientific data, however. Furthermore, though the collaboratory model does require coordination across research groups, it does not necessarily entail collaboration among them.

We are working on developing yet another model for the ownership and use of scientific data. It is a model we term the data collective. Unlike data corpora designed to provide shared access to basic data for researchers within a discipline, we propose a model that involves sharing of both basic data and intermediate findings among researchers from different disciplinary traditions. A prototype of this sort of collaboration can be seen in the Professional Competency Project. This project was initiated to foster multidisciplinary research on clinical problem-solving and the assessment of medical competency. The medical licensing authorities in the U.S. and Canada have implemented (or are in the process of implementing) performance-based assessment for all licensure candidates. Such forms of assessment involve working up clinical cases presented by "standardized patients" (SPs), that is actresses or actors trained to portray a testing case. The National Board of Medical Examiners (NBME), for example, has announced its intention to add a test of clinical skills using SPs to the United States Medical Licensing Examination (USMLE) by the fall of 2004 (NBME, n.d.). Standardized patients have been used as a part of the certification process employed by the Educational Commission for Foreign Medical Graduates since 1998(ECFMG, 2002) and such tests have been used in the licensure exams conducted by the Medical Council of Canada (MCC, 2002) for over a decade. Despite its importance in the training, licensure, and certification of medical practitioners, performance based assessment has been little studied outside of the institutions within which it is used because of the difficulty and expense involved in recruiting subjects, the lack of access to adequately-equipped testing facilities, and the unavailability of useable testing cases. The Professional Competency Project involves producing a corpus of performance-based assessment protocols for study by a diverse collection of social scientists.

We are in the process of developing a set of testing cases based on real clinical data. Research subjects will be undergraduate medical students and residents

enrolled at a particular medical school. All subjects will be asked to take a history and perform a physical exam on SPs trained to present the testing cases in a special testing facility known as the Professional Development Laboratory (PDL). The PDL consists of a suite of rooms equipped as clinical examination rooms. Each examination room in the PDL contains all of the paraphernalia (e.g., exam table, otoscope, sphygmomanometer, eye chart) normally used in conducting a physical examination. After working up the patient, subjects retire to an adjacent computer lab and are given 30 min to compose a SOAP note and order lab tests. A SOAP note is a structured chart entry consisting of the subjective data (i.e., the patient's reported symptoms), objective data (i.e., physical findings and lab results), assessment (i.e., a differential diagnosis), and a treatment plan. After completing their SOAP note and lab orders, the subjects were taken to an interview room in which a structured debriefing is conducted. The results of the laboratory tests ordered by the subject are presented to the subjects in the context of this interview. The videotapes from the encounters with the SP and the debriefing interviews are transcribed by a medical transcriptionist. These protocols will be studied by a group of researchers at different institutions to address different research questions.

One foundational issue pertains to the forms of knowledge subjects mobilize and employ in clinical problem solving. One approach to addressing this issue is to conduct a "cognitive discourse analysis" (Frederiksen, 2001). Cognitive discourse analysis (CDA) begins by constructing models of experts' procedural and declarative knowledge in particular domains of problem solving. The propositional content of subjects' discourse contributions can then be mapped with reference to the expert task model so constructed as a means of diagnosing possible deficiencies in the subjects' preparation or understanding.

A host of related questions can be raised with respect to the ways in which subjects reason through diagnostic problems. How do they initially come to formulate the problem, for instance? How do they amend their understanding of the patient's problem as new information is developed? How is conflicting data dealt with? Such questions are related to the way in which the subject organizes the problem in memory. An analysis of memory-based reasoning can reveal past principles and experiences stored in memory, circumstances providing access to information, competing solutions, problem features used to test whether information is applicable, and support for the solution selected by examining subjects' problem-solving protocols (Seifert, Patalano, Hammond, & Converse, 1997). Performance-based assessment requires that the subject and SP work together to produce a "case". Similarly, the purpose of the post-encounter interview is to reveal the subject's reasoning, but, again, the thing that the participants orient to (the "case") is something produced interactionally. One might ask what methods do these members (subject, SP, interviewer) employ in carrying out this interactional task? The discipline that has taken up such matters as its central area of concern is Ethnomethodology (Garfinkel, 2002; Heritage, 1984). Conversation Analysis (CA)

is an area of specialization within ethnomethodological research that focuses on the methods members use in interaction to make sense of each other and to, in turn, be seen and heard as sensible (Psathas, 1995; Heritage, 1984). Cicourel (1975) studied how interaction between a doctor and patient is transformed in producing a chart note. The corpus produced in this project constitutes a unique set of materials for studying the same process, one in which the "case" is not only reproduced in the chart note, but also in an elaborate post-encounter interview.

The encounter between the SP and the subject results in two records, the subject's SOAP notes and the debriefing interview. Recent developments in computational linguistics have provided ways to statistically represent the semantics of these records allowing for a comparison of content. Latent semantic analysis (LSA) is one such method. For instance, these records can be compared with a gold-standard provided by an expert. Such a comparison has been successfully used in automated essay grading, where a student essay is compared with an expert essay in order to measure the student's performance (Landauer, Foltz, & Laham, 1998). LSA has generally been used for natural texts, however. Whether it can equally successfully applied to structured set of chart notes and dialog remains an important question to be investigated.

The four types of analyses (cognitive discourse analysis, analysis of memory-based reasoning, conversation analysis of doctor/patient interaction, latent semantic analysis) arise out of specific interests and address particular questions. They are not completely independent, however, and in many cases individual researchers can benefit from and build upon the analytic findings of others. Cognitive discourse analyses and analyses of memory-based reasoning, for instance, have considerable overlap in scope. The task model for a particular case developed to support cognitive discourse analysis would also be of great value for conducting analyses of memory-based reasoning. A detailed analysis of the reasoning done by a subject for a particular case would doubtless also be useful to an analyst doing a cognitive discourse analysis of the debriefing interview for that subject. Enhancements to the working transcripts produced by the CA researchers may prove to be useful to researchers doing analyses of the propositional content of dialog or memory-based reasoning. The more global measures of comprehension generated using LSA will be of value for all researchers in identifying particular assessment protocols in which problems in understanding were evidenced. The goal, therefore, is to go beyond the forms of sharing of "final research data" mandated by federal funding agencies (e.g., NIH, 2003) to include findings and other improvements produced by collaborating researchers. A number of questions arise, however, with respect to the construction of a data collective such as the one described here. When the research involves human subjects (as it does here), what are the ramifications of the data collective model with regard to confidentiality and informed consent? Also, what kinds of institutional arrangements need to be put into place to protect the interests of the researchers who contribute their work to the project? Finally, what are the

requirements from a technological perspective for supporting this form of collaboration?

References

Cicourel, A. (1975). Discourse and text: Cognitive and linguistic processes in studies of social structure. Versus, 12(11), 33‹84.

ECFMG (2002). Clinical skills assessment: Candidate orientation manual. Philadelphia, PA: Educational Commission for Foreign Medical Graduates. Retrieved August 28, 2002 as http://www.ecfmg.org/csa/csacom.pdf

Frederiksen, C.H. (2001). Propositional representation in cognitive psychology. In N.J. Smalser &P.B. Baltes (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Amsterdam: Elsevier Science.

FSMB (n.d.). Position of the Federation of State Medical Boards in support of adding a clinical skills examination using standardized patients to the United States Medical Licensing Examination (USMLE). Retrieved August 29, 2002 ashttp://www.fsmb.org/Policy%20Documents%20and%20White%20Papers/standardized_patient _support_white_paper.htm

Garfinkel, H. (2002). Ethnomethodology's program: Working out Durkheim's aphorism. Lanham, MD: Rowman & Littlefield Publishers.

Heritage, J. (1984). Garfinkel and Ethnomethodology. Cambridge: Polity Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.

MacWhinney, B. (1995). The CHILDES project: Tools for analyzing talk (2nd Ed). Hillsdale, NJ: Lawrence Erlbaum Associates.

NBME (n.d.). USMLE clinical skills examination. Retrieved August 28, 2002 fromhttp://www.usmle.org/news/newscse.htm

NIH (2003). NIH data sharing policy and implementation guidance. Bethesda, MD: Office of Extramural Research, National Institutes of Health. Retrieved May 30, 2003 as: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Psathas, G. (1995). Conversation analysis: The study of talk-in-interaction. Thousand Oaks, CA: Sage.

Seifert, C. M., Patalano, A. L., Hammond, K. J., & Converse, T. M. (1997). Experience and expertise: The role of memory in planning for opportunities. In P. J. Feltovich, K. M. Ford & R. R. Hoffman (Eds.), Expertise in Context: Human and Machine (pp. 101 - 123). Menlo Park, CA: AAAI Press/ MIT Press.

Organizational Changes and Learning in a Laboratory

Flemming Meier
Danish University of Education, Denmark

This is a position paper for the workshop on Computer Supported Scientific Collaboration (CSSC). The aim of the paper is to present empirical ethnographic fieldwork in a cellbiological laboratory carried out this spring (2003) as part of a ph.d.-project. The paper will start out with a description of the ph.d.-project and then turn to a presentation of the empirical studies - focusing on the goals, the methods, some of the research questions and a brief presentation of some preliminary results. In the end a couple of questions will be outlined - questions that the present state of the project and the results from the investigations in the laboratory seem to press forward.

The project

The overall aim of the ph.d.-project is to contribute to the development of new ways of studying and understanding organizations, organizational change(s) and organizational learning.

More specifically the goal of the project is to deal with the following issues and questions:

- Development of methods to do empirical investigations and describe organizations and organizational changes. If an organization is something more than the aggregation of individuals - a kind of a superorganism - what are then important matters and things to focus on in empirical investigations?
- Investigation and analysis of questions of exchange, development and institutionalisation (or 'technologization') of knowledge in organizations. How is knowledge 'build' into the organization? How is certain knowledge

'remembered' by the organization as individual persons 'pass through' the organization?
- Investigation and analysis of the role of technological systems and artifacts in organizational change / learning processes.

The term organizational changes cover both small 'unnoticed' changes in the everyday life of the organization and more comprehensive (often) planned changes / restructurations. Further both structural and cultural changes are relevant. Learning is perceived as complex processes of action and participation - involving individual as well as social or collective dimensions - and thus having both personal as well as organizational, cultural and other aspects. This view also stresses the contextuality of learning because learning is seen as closely connected to concrete activities or practices embedded in social and material contexts Both learning processes of individual persons, groups of persons ('in' the organization) and the learning processes of the organization are relevant. However a certain focus is given to the latter. A basic question here is probably whether it is wise at all to ascribe 'learning abilities' to organizations. Do organizations learn or do they just change?

From February to end of May 2003 - ethnographic investigations in a cellbiological laboratory have been carried out. The plan is to do empirical work in another organization in the end of 2003. Then there will be a lot of analytical work, papers to be written, seminars to be attended etc. The project will result in a dissertation to be delivered august 2005.

The fieldwork

In line with goals of the project, the goals of the empirical studies were to:
- Test methods
- Produce insights in knowledge construction + exchange and institutionalization (authorization) of knowledge
- Gather data about technology and the role of technology in the constitution and change of the lab as an organization

The investigations have consisted of:
- Observations on a broad scale on what is going on in the laboratory. A great number of work processes have been observed and recorded (descriptions in text and small video-clips). This has included both observations of experimental laboratory work, laboratory meetings, employees attending external meetings and conferences, other meetings or conversations between employees and other events and work processes in the laboratory.
- Informal as well as more formal and structured interviews about the research projects, the work in general and the laboratory as an organization / workplace. A small number of structured interviews with individual

employees and many informal interviews in close connection with observations have been conducted.

- A systematic registration of technological and other artifacts in the laboratory. Including for instance pipettes, machines, laboratory benches, chairs, computers, cancer cells, incubators, freezers, closets, containers, project log-books, lists of cell lines, books, manuals, assays, protocols, lab coats, posters, tables, microscopes, printers aso. Including a registration of function, relations or connections out of the laboratory / the organization, who uses the artifact (and who do not), technological history (if it is possible or relevant to speak of one) and descriptions of various situations of use.
- Close observations of daily practice in two selected research projects, focusing on the connections or 'touch points' between the projects and the influence such connections have on the directions and developments of the projects in terms of inspirations between persons and projects when it comes to conceptualization of research problems / questions, experimental methods or techniques and overview of the research field..

Some of the more specific or empirical questions were:

- What organizational changes are going on in the organization in question? How can the changes be categorized and described? What learning processes emerge as 'answers' to organizational changes or as utilizations of new possibilities for action and participation? Can the learning processes be categorized in terms of (for example) reproductive / innovative and individual / collective / organizational? How are needs for learning articulated in the atmosphere of organizational change? Are there any connection between certain types of organizational change and certain types of learning processes? Do certain types of organizational change provide certain conditions for (or hinder) learning processes?
- Knowledge - Which forms and dimensions of knowledge exist and are 'applied' and expressed in the everyday life of the organization and the work practices? How are various forms or patterns of constructing and sharing of knowledge changed? What is the role of ICT-media, e-learning systems and other uses of computers in constructing and sharing of knowledge?
- The role of technological systems in relation to learning processes. Is it possible to formulate a useful typology of technological artifacts for analysing the role such artifacts play - for example as a resourceful, legitimising, hindering, opening, excluding factor - in learning processes in a workplace / organization?

The laboratory

The laboratory is called the Apoptosis Department. It is part of the research section of the Danish Cancer Society. The laboratory has existed for about 10 years now - first as a small group of researchers in another department, then later with the status of a formal subgroup and recently with the status as a department. Currently around 15 people are employed in the laboratory. A leader, four post-docs, four ph.d.-students (or post-grads about to formulate a ph.d.-project), three masters students and two laboratory technicians.

The laboratory conduct research on programmed cell death - also called apoptosis. The aim of the research is to understand the sequence of invents in (cancer) cells that lead to apoptosis by identifying proteins that inhibit this pathway and by studying the mechanisms by which they do so.

The research work is organized around projects. Each employee - except the lab techs - has their own research project with specific goals and funding. Out of the projects the work tasks - and occasionally collaborations - evolve. Each week there is a lab meeting where all kinds of practical things are discussed as well as the concrete research problems. In connection to these meetings each researcher in turn also give presentations on their projects. There are also (spontaneously as well as long time planned) meetings between students and supervisors. The leader and the post-docs are supervisors for the others, and the leader is supervisor for the post-docs.

Preliminary results of fieldwork

A brief initial evaluation of methods in the ethnographic studies show that:

- Video clips captures mostly body movements in contexts – good for work operation analysis - it provides a rich picture of work processes and knowledge forms - but few clues / hints of organizational matters
- Systematic registration of tech artifacts produces a lot of redundant data - but it also provides a deeper insight in laboratory / research techniques
- Descriptions of observations capture data about organizational culture.

The kinds of organizational changes that were recorded vary from the change of status of the laboratory (from group to department) to a tightening of sterility procedures and many small changes in localization of things and what-to-do in certain situations ("when you take the second last assay inform the lab tech"). Small changes are negotiated in an interplay between lab work situations (the informal talk) and lab meetings (the formal decision).

Very many forms of knowledge were recorded. Examples are:

- Many tiny pieces of 'where is this and that and what to do with things'
- Knowledge of how to operate certain instruments and machines
- Knowledge about a wide range of techniques
- Knowledge about some other peoples projects and the current state of these
- Scientific knowledge about field
- Different degrees of overview of the various projects and how they interact and contribute to the overall goals of the laboratory or research field
- Collections of data
- Lists of things – attributes, locations
- Descriptions in lab-logs
- Instructions and descriptions in protocols
- Knowledge in articles, books etc.
- Info on posters
- Knowledge 'build' into things

The analytical task is now to sort these and other examples in order to answer the research questions mentioned above.

On the role of technology at least three things seem to be important:

- Technology is highly integrated in the work
- Tweaking of techniques, 'alternative' use of tech and varying combinations of techniques and instruments are essential for the experimental and innovative work
- Many 'small' collaborations and learning situations evolve around the use of a certain technique, instrument or machine

It should be stressed that these results are preliminary, and that a thorough going through the data will probably lead to reformulations and additional results.

Questions

Two issues or questions that has grown out of the work (empirical as well as analytical):

- The interplay between the individual projects and the constitution of a 'wholeness' of projects.. How is the projects as a whole overviewed in the organization - and by whom? How does interplays, connections, collaborations and other 'contact points' happen? What are the implications of such 'contact points' for further directions of the projects?
- Challenges and demands from collaborators and competitors (outside the laboratory). What kinds of challenges etc.? How are challenges, expectations,

demands etc. perceived - and by whom? How does people or the organization react?

These are questions that the further work might focus more on. The laboratory will be visited again this fall in a period of 3-4 weeks. Mainly structured interviews with selected employees in the lab will be conducted.

How Outsourcing Impacts to Decision-Making over IS Process Innovations

Erja Mustonen-Ollila
Lappeenranta University of Technology, Department of Information Technology, Finland

Abstract. This paper examines how outsourcing impacts to decision-making over Information System (IS) process innovations (ISPIs) in three organisational environments over four decades using a sample of 77 decision-making events. In the analysis the four decades are divided into four time generations: 1) 1954-1965, 2) 1965-1983, 3) 1983-1991, and 4) 1991-1997. These follow roughly Friedman's and Cornford's (1989) categorisation of IS development eras. Four types of ISPIs are distinguished: base line technologies (T), development tools (TO), description methods (D), and managerial process innovations (M). Three types of decision-maker groups are found: IS department and IS client; IT department, IS client, and IS vendor; and IT department, IS client, and two IS vendors. The analysis shows, that before outsourcing in 1984 decision-making was centralised between the IS client and the internal IS department. After outsourcing decision-making became distributed, where the IS client, the two IS vendors, and the internal IT department decided over ISPIs.

Introduction

Outsourcing is the turning over of an IS, in whole or in part, to one or more external service providers in order to supply human or technical IS resources to the organisation (Soininen, 1995, 1997; Gray, 1994). It also implies the use of external agents to perform an organisational activity (King and Malhotra, 2000). Two of the main reasons for outsourcing are the lack of resources, and to have access to emerging technologies that have the potential to change fundamentally the firm's

business processes (McLellan and Marcolin, 1994). Once IS is outsourced, top management no longer has direct command authority over it (King, 1994; McLellan and Marcolin, 1994). IS project management from the point of view of the service receiver is now being carried out by people not under their supervision (Saarinen et al., 1995; Grover and Teng, 1993). Supplier group can be dominated by few companies, particularly when the client is itself a large organisation (King and Malhotra, 2000; Lowell, 1992; McFarlan and Nolan, 1995; Foxman, 1994; Meyer, 1994). The contract with the vendor is important as to ensure that expectations are met (Lacity and Hirschheim, 1993a, 1993b; Gray, 1994; Lacity et al., 1995). If the executives want to gain independence concerning IS development, they may want to decrease the influence of the centralised IS department by outsourcing (Lacity and Hirschheim, 1993a, 1993b). Pfeffer (1981), and Lacity and Hirschheim (1993a, 1993b) suggest that to understand decisions one should focus on the power of the IS department, the vested interests of different decision-maker groups, and the political tactics they may enact to sway decisions in their favour.

Information System Process Innovations (ISPIs) have become important for organisational effectiveness (Swanson, 1994), and we presume that a specific ISPI is chosen for use at a specific ISD project (Rogers, 1995). ISPIs cover technological, organisational or administrative innovations, and we classify ISPIs into four categories (Mustonen-Ollila and Lyytinen, 2003a, 2003b). Furthermore, based on Friedman and Cornford (1989) we classify ISPIs into four eras (Mustonen-Ollila and Lyytinen, 2003a, 2003b): the first generation from the late 1940s until the mid 1960s; the second generation from the mid 1960s until the early 1980s; the third generation from early 1980s to the beginning of 1990s; and the fourth generation from the beginning of 1990s. Our main research question concerning ISPIs in the context of outsourcing was "*If and how decision-making over ISPIs as a result of outsourcing has favoured or unfavoured the different decision-maker groups?"*

Research Methodology and Findings

We chose a qualitative case study (Laudon, 1989; Johnson, 1975; Curtis et al., 1988) with a multi-site study approach, were we investigated three Finnish organisational environments, known here as companies A, B and C, respectively. Company A is a big paper manufacturer, whereas company B and C are specialised in designing, implementing, and maintaining information systems. Our study forms a descriptive case study (Yin, 1993): it embodies time, history and context, and it can be accordingly described as a longitudinal case study, which involves multiple time points (Pettigrew, 1985, 1989, 1990). Research approach followed Friedman's and Cornford's (1989) study, which involved several generations and time points. Because the bulk of the gathered data was qualitative, consisting of interviews and archival material, we adopted largely historical research methods (Copeland and McKenney, 1988; Mason et al., 1997a, 1997b). Our definitions of outsourcing

issues over ISPIs formed the basis for interviews and the collection of archival material (Järvenpää, 1991). We used triangulation by checking simultaneously different data sources, such as the archival material to improve the reliability and validity of the data. At the data categorisation stage the internal IS department, the CAIS department, the IS client, and the two IS vendors were classified into three decision-maker groups. The decision-maker groups were further divided into four time generations, and four ISPI categories. Thereafter, 77 decision-making events over ISPIs were found, and these events within the each decision-maker group over time were summed up (See table 1). We excluded time generation one, because it lacked data.

Decision-maker Groups one, two, and three	**Time Generation**	**M**	**T**	**TO**	**D**	**Total Number of Decision-making Events over ISPIs**
One: IS department and IS client	Two (1965-1983)	17	15	6	5	**43**
Two: CAIS department, IS client, and first IS vendor	Three (1983-1989)	1	5	6	3	**15**
Three: CAIS department, IS client, first IS vendor, and second IS vendor	Four (1989-1997)	1	10	6	2	**19**
		19	30	18	10	**77**

Table 1. The number of the decision-making events over ISPIs.

Data in table 1 was analysed using the chi-square test (Brandt, 1976). The test showed statistically significant differences (χ^2=14,609, α=0.05) which shows that in different decision-maker groups decision-making in ISPI groups vary dramatically over time.

Discussion and Conclusions

In this study we analyse the longitudinal data sets of decision-making over ISPIs in three decision-maker groups over time. The results show that until 1984 outsourcing decision-making was centralised and the internal IS department was very influential over ISPIs. After 1984 outsourcing making became distributed between the CAIS department, the two IS vendors, and the IS client. The CAIS department became interested in buying IS solutions, and employed system managers who became influential decision-makers over ISPIs. Before nominating as system managers, they had acted as the main end-users in a specific IS business area. Their responsibility as system managers in the IS projects was to gather the needs of the end-users, and to introduce the needs to the IS vendors. The system managers made co-operation with the IS client and the IS vendors. The CAIS department consulted with the personnel administration's, and the forest department's IT departments. The project managers of the IS client and the IS vendors', and the CAIS department's system managers decided mutually over ISPIs. IS project working in the factory systems, in

many IS project environments, and in matrices organisations caused a lot of problems, and co-operators were needed.

When comparing our results to the related research we were able to strongly ascertain a need for co-ordinating persons (King and Malhotra, 2000; McFarlan and Nolan, 1995; Foxman, 1994; and Meyer, 1994), and IS client must control its own business tied to Information Systems, when it uses IS vendor in implementation (Lowell, 1992). No support was found to loss of strategic control over the application of IT resources, and that IS client no longer controls IS project work (Saarinen et al., 1995; Grover and Teng, 1993; Lacity et al., 1995). It was also confirmed that the IS client wanted to decrease the political power of the internal IS department by making an outsourcing decision (Lacity and Hirschheim, 1993a, 1993b). The CAIS department's role as a mutual co-ordinator between many interest groups became vital, because it had specific IS business area knowledge. After 1984 company level information systems were sales and order handling systems, and accounting systems. The CAIS department was responsible for all these information systems. The CAIS department had several other responsibilities as a co-ordinator, such as consulting the personnel administration's IT department, the forest department's IT department, and the factories about ISPIs and the mutual working procedures. It worked in the IS project steering groups with the IS vendors, bought IS solutions, and technology. The CAIS influenced directly the company level Information Systems, it made the IS project contracts, and negotiated with the IS vendors. After 1984 the IS client wanted to buy Information Systems from the IS vendors in a fair price, and it wanted to co-operate with the IS vendors to ensure that Information Systems would serve the business in the company, but also to support the IS vendors to gain other clients. The IS vendors on the other hand were obliged to learn IS business knowledge of the IS client.

References

Brandt, S. (1976) Statistical and Computational Methods in Data Analysis, North-Holland Publishing Company.

Copeland, D.G., McKenney, J.L. (1988) Airline Reservations Systems: Lessons from History, MISQ, pp. 353-370.

Curtis, B. Krasner, H., Iscoe, N. (1988) A Field Study of the software design process for large systems, Communications of the ACM, 31 (11), 1268-1287.

Foxman, N. (1994) Succeeding in Outsourcing: Cultivate the Outsourcing Relationship, Information Systems Management, pp. 77-80.

Friedman, A., Cornford, D. (1989) Computer Systems Development: History, Organization and Implementation, John Wiley & sons.

Gray, P. (1994) Outsourcing and Other Strategies, Information Systems Management, 11, pp. 72-75.

Grover, V., Teng, T.C. (1993) The Decision to Outsource Information Systems Functions, Journal of Systems Management, 44, pp. 34-38.

Johnson, J.M. (1975) Doing field research, The Free Press.

Järvenpää, S. (1991) Panning for Gold in Information Systems Research: 'Second-hand' data, Proceedings of the IFIP TC/WG 8.2, Copenhagen, Denmark, December 14-16, pp. 63-80.

King, W. (1994) Strategic Outsourcing Decisions, Information Systems Management, pp. 58-61.

King, W.R., Malhotra, Y. (2000) Developing a framework for analyzing IS sourcing, Information & Management, 37, The International Journal of Information Systems Applications, pp. 323-334.

Lacity, M.C., Hirschheim, R. (1993a) The Information Systems Outsourcing Bandwagon, Sloan Management Review, 35, pp. 73-86.

Lacity, M.C., Hirschheim, R.(1993b) Information Systems Outsourcing- Myths, metaphors, and realities. John Wiley and Sons.

Lacity, M.C., Willcocks, L.P., Feeny, D.F. (1995) IT Outsourcing: Maximize Flexibility and Control, Harvard Business Review, pp. 84-93.

Laudon, K.C. (1989) Design Guidelines for Choices Involving Time in Qualitative Research, Harvard Business School Research Colloquium, The Information Systems Research Challenger: Qualitative Research Methods, 1, pp. 1-12.

Lowell, M. (1992) Managing Your Outsourcing Vendor In The Financial Services Industry, Journal of Systems Management, 43, pp. 23-36.

Mason, R.O., McKinney, J.L., Copeland, D.C. (1997a) Developing a Historical Tradition in MIS Research, MIS Quarterly, 21 (3), 257-278.

Mason, R.O., McKenney, J.L., Copeland, D.G. (1997b) Developing a Historical Tradition in MIS Research, MIS Quaterly, 21 (3), 307-320.

McFarlan, F.W., Nolan, R.L. (1995) How to Manage an IT Outsourcing Alliance, Sloan Management Review, 36, pp. 9-23.

McLellan, K., Marcolin, B. (1994) Information Technology Outsourcing, Business Quaterly, 59, pp. 95-104.

Meyer, N.D. (1994) A Sensible Approach to Outsourcing, Information Systems Management, pp. 23-27.

Mustonen-Ollila, E., Lyytinen, K. (2003a) Why organizations adopt IS process innovations: A longitudinal study using Diffusion of Innovation theory. Information Systems Journal, 13 (in press).

Mustonen-Ollila, E., Lyytinen, K. (2003b) How Organisations adopt Information System Process Innovations: A Longitudinal Analysis. Accepted for publication for EJIS.

Pettigrew, A. (1985) The Wakening Giant, Continuity and Change in ICI.

Pettigrew, A. (1989) Issues of Time and Site Selection in Longitudinal Research on Change, The Information Systems Research Challenger: Qualitative Research Methods, 1, pp. 13-19.

Pettigrew, A.M. (1990) Longitudinal field research on change: theory and practice, Organization Science, 1 (3), 267-292.

Pfeffer, J. (1981) Power in Organisations, Pitman Publishing, Marchfield, Massachusetts.

Rogers, E.M. (1995) Diffusion of Innovations, Fourth Edition, The Free Press.

Saarinen, T. Salmela, T., Vepsäläinen, A.P.J. (1995) Outsourcing of Information Systems Services in Finnish Companies, Helsinki.

Soininen, J. (1995a) The Dynamics of IS Outsourcing. Available in the Internet http://www.ocuf.fi/~soininen/os/out1.html.

Soininen, J. (1997) A Framework for Analyzing of Outsourcing in the Information Technology field, M.Sc. thesis, University of Waterloo, Ontario, Canada.

Swanson, E. B. (1994) Information Systems Innovation Among Organizations, Management Science, 40 (9), 1069-1088.

Yin, R.Y. (1993) Applications of Case Study Research, Applied Social Research Methods series, 34, SAGE publications.

Decentralized Knowledge Discovery for Scientific Collaboration

Giuseppe Psaila and Davide Brugali
Università degli Studi di Bergamo, Facoltà di Ingegneria, Italy

Abstract. Knowledge discovery processes require powerful computational resources, and specific expertise to extract knowledge from large amounts of data. In the context of scientific collaboration, such as a Network of Excellence to which world-wide scientific partners participate, each partner provides some data resources or some computational resources or some expertise. Thus, decentralization of knowledge discovery processes seems a viable solution. However, this kind of decentralization can be effective only if there is a common framework within which the process is carried on and all resources are integrated and shared. In this paper, we illustrate our ideas about *Decentralized Knowledge Discovery for Scientific Collaboration*. Primarily, we discuss technical issues concerning the decentralized execution of knowledge discovery activities. Nevertheless, we also discuss the (positive, we hope) social impact of the proposed technical solution.

Introduction

The Internet connects people, resources and activities. It facilitates the exchange of information and supports the co-operative work of managers, analysts, engineers, etc. This is true for scientific collaboration contexts as well. Consider a *Network of Excellence* composed by world-wide partners, such as scientific institutes and research centers, which collaborate to carry on a common research. Some of them operate on the application field of research and with end users; some of them provide computational resources; some of them provide expertise. For example, a network working on human diseases may involve hospitals (which collects data), physicians (the end users, which expect new therapies), biologists (domain experts)

and specialized research centers (which own the computational and human resources to perform complex data analysis).

Data mining and knowledge discovery activities usually play an important role, since scientific activities are strongly based on data gathering and analysis. Partners of a Network of Excellence brings their own specialized capabilities and resources (computational and human) to these activities. The integration, through the internet, of computational and human resources may give significant benefits, but the knowledge discovery process becomes necessarily *decentralized*, since it involves decentralized resources.

In this paper, we illustrate our ideas about *Decentralized Knowledge Discovery for Scientific Collaboration*. We try to understand which are the basic elements that constitute a Knowledge Discovery Process (KDP) performed in a decentralized way. We discuss how a common framework for decentralized knowledge discovery processes may be effective to carry on complex knowledge discovery processes, in particular in terms of social impact that we expect to be positive in terms of collaboration improvement.

Actors in Decentralized Knowledge Discovery for Scientific Collaboration

Scientific collaboration is based on the cooperation of several partners, usually institutes and research centers, as in the case of the Network of Excellence. Each partner is specialized in conducting particular activities, usually related to knowledge discovery tasks. To carry on scientific collaboration, partners typically share data, expertise, computational resources; this means that the process is totally or partially moved through these resources.

Let us understand what kind of actors are necessarily involved in such a process.

Data Holders. Data holders are those who actually hold the data to analyze. In the network of excellence, some partner may be specialized in data gathering from the research field.

End Users. End users are those who wish to take advantage of the (unexpected) knowledge that can be discovered from within the data. In the Network of Excellence, end users are those that benefits from the results (for examples, physicians which receive information about diseases and about new therapies).

Hardware/Software Holders. This category includes the owners of well equipped computational resources and KDD software tools specialized for data mining and knowledge discovery. In the case of Network of Excellence, these resources might be hold by a subset of the partners, those specialized in data analysis activities (e.g. centers specialized in super-computing).

Analysts. Computational resources and knowledge discovery tools are useless without skilled human resources that are able to exploit them in the analysis process. To do that, it may be necessary to exploit both experts in the specific application

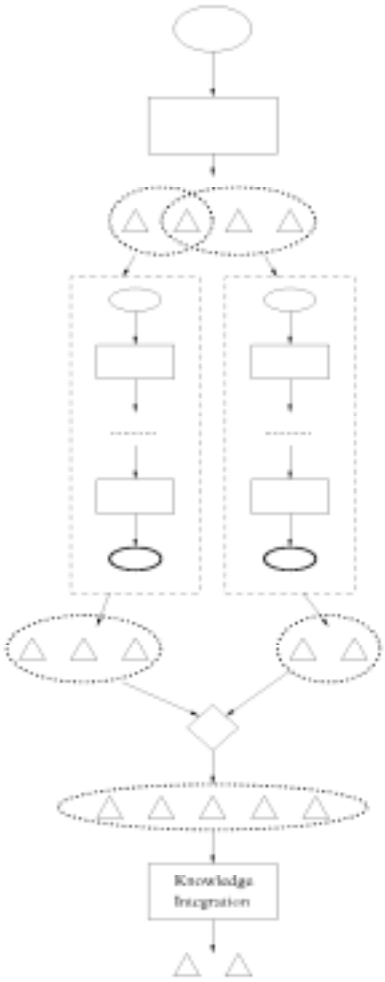

Figure 1: Dynamic Workflow

domain (for instance, experts in biology) and experts in the conduction of knowledge discovery processes (these are the technicians which are able to interpret requirements in order to adopt the proper knowledge discovery techniques and tools).

Decentralized Knowledge Discovery Process

Let us discuss our ideas about the fundamental activities for decentralized KDPs, involving the identified actors.

Activities. At first, let us discuss the set of activities which constitute the decentralized knowledge discovery process.

Data Gathering. In the decentralized scenarios, it is necessary to identify source data sets involved in the process. During this activity, data are actually collected (to

constitute the initial database of the process, or to update old data sets with new data sets), or simply made accessible and linked to the process.

Selection. The selection activity chooses the data on which a specific subtask of the KDP is focused, among all available data sets. In fact, the KDP might be composed of several subtasks, each one working on a different subset of the collected data sets and performed by different groups.

Preprocessing. The preprocessing phase is necessary to remove noise and incomplete data from the data sets selected in the selection activity, and involves application domain experts and data cleaning experts.

Simplification. The simplification activity consists in simplifying and transforming the data set to analyze, in order to make it suitable for the chosen data mining tool. Here the expert of data mining tasks has the keys of the activity.

Data Mining. Data mining and data analysis tools are actually applied to selected and simplified data. It is necessary to define the parameters that drive the data mining tool, so that the generated models (patterns) are significant.

Evaluation. If the extracted models/patterns are not considered useful or they are considered not accurate by domain experts, the reasons why such patterns are not adequate should suggest how to modify the parameters featuring the execution of the previous activities. Otherwise, the generated models/patterns are made available for next activities; they are pieces of knowledge for the overall process.

Knowledge Integration. An activity focused on knowledge integration is fundamental. In fact, pieces of knowledge may be separately discovered by different teams involved in the process, but in order to achieve a full comprehension of the overall studied phenomena it is necessary to coherently integrate these pieces of knowledge. The result is the knowledge base of the overall process (that can be updated several times during the process).

Knowledge Delivery. Finally, the knowledge is delivered to end users. Observe that this is not a trivial task, since not necessarily the complete knowledge base is of interest for end users (for example, physicians are informed about effectiveness of new therapies). Observe that this activity may produce another result: it may happens that new, previously unexpected, needs or ideas come out, suggesting to perform other knowledge discovery activities.

Meetings are another kind of activity, which are usually important in the KDP, since they are the occasion to discuss the results and to make decisions. The organization of classical meetings is not trivial in the scientific collaboration context, due to people which have to move to the meeting venue: virtual meetings are a better solutions. Observe that the results of meetings (reports, decisions, etc.) might be significant to move on the KDP, therefore they must be explicitly considered as a (special) activity for the decentralized knowledge discovery process.

Tasks and Subtasks. A Task is a sequence of knowledge discovery activities, performed with a specific goal; a task can contain specific subtask. Each task or subtask has a supervisor, which is responsible to move on the assigned task. This

way, the knowledge discovery process is partitioned into possibly parallel processes, that can be performed by different teams on different hosts. At the end, the supervisor of the main task collects the results, integrates and delivers the knowledge, possibly creates new subtasks.

Parallelism and Decentralization. Both the topics are strictly correlated, since decentralization implies, in some sense, parallelism. In our context, knowledge discovery activities are necessarily decentralized, since they are delegated to each research center participating to the network; this means that each center is responsible to carry on specific activities, even complex, and therefore better performed inside the center (this fact implies mobility of activities and tasks). However, there is no need that decentralized activities are necessarily sequential, but often they can be executed in parallel (e.g. on different hosts).

Observe that decentralization implies resource distribution and resource mobility. However, technical issues concerned with mobility are outside the scope of this paper.

Workflow Support

The decentralized KDP, discussed in the previous section, is based on concepts that are typical of *workflow models*. In effect, we can imagine the KDP as a special kind of workflow, in which the sequence of activities and subtasks is dynamically built. Note that traditional workflow concepts are not suitable for our context: in fact, the sequence of activities to perform in knowledge discovery strongly depends on the partial results, thus it must be defined dynamically. We launch the idea of *Dynamic Workflow* for KDP, i.e. a workflow where structure of processes is dynamically built. Obviously, a software for supporting KDP based on dynamic workflow is crucial to make all that feasible.

We refer to Figure 1, which reports a possible graphical representation for a sample KDP process modeled as a dynamic workflow. Symbols in the figure have the following meanings. *Thin ovals* denotes start symbols for tasks and subtasks, while *thick ovals* denotes stop symbols. *Solid-line rectangles* represent knowledge discovery activities, while dashed-line rectangles denote subtasks (inside their, we can find again start and stop symbols, activities, etc.). Triangles denote data sets, which are generated by activities and subtasks; dotted ovals denote groups of data sets. Finally, *diamonds* represent convergence symbols, in which parallel activities or subtasks are synchronized.

The sample workflow denotes an on-going knowledge discovery task, since the stop symbol is not present; this means that a user, with the role of task supervisor, may decide to define new activities and subtasks; for each of them, the task supervisor defines a set of requirements (to instruct people involved in the activity/subtask to properly perform it), defines the set of input data sets, assigns activities and subtasks to the proper working team.

At the beginning, subtasks are empty; they must be defined by the person of the working team to which the subtask has been assigned, i.e. the subtask manager; similarly, for activities as well it is necessary to define the activity manager, which has the responsibility to carry on the activity. During the execution of single activities, the working team may exploit any kind of knowledge discovery tool suitable for the specific type of activity.

Subtasks and activities may be executed in parallel. For instance, this is the case of the two subtasks reported in the figure. When they finished, each of them produced a pool of data sets. Then, the general task is synchronized (diamond symbol) and all the data sets produced by the subtasks are made available for the whole main task.

Finally, all of the data sets are used by the *Knowledge Integration* activity, which is responsible to integrate pieces of knowledge discovered by the two independent subtasks, generating new data sets which may constitute a first result of the knowledge discovery task (i.e. knowledge).

Social Issues

We expect that the adoption of a system based on the decentralized knowledge discovery framework will have significant and positive (we hope) social impact. In particular, we consider the following aspects: improvement of collaboration, ease of knowing how the collaboration is going on, personal satisfaction and trust about data. Let us discuss these points in detail.

Improvement of collaboration. Scientific collaboration usually involves research centers which are located far away each other. The distance is a factor that usually reduce the degree of collaboration. In fact traveling is very expensive, in terms of money and time. This situation causes dissatisfaction, since if people do not travel, the collaboration cannot be carried on effectively, but if people travel too much, they are unsatisfied as well (traveling is tiring and distracts from working on research issues). The situation is even worse if we consider that it is difficult to organize meetings in dates which are good for many people.

These considerations motivate the idea of virtual meetings, i.e. meetings organized though the Internet by means of computers, cameras, microphones. They are easy to organize and certainly less expensive than travels. We expect significant improvements as far as the collaboration is concerned: every time the need arise, a virtual meeting can be organized; this way, the collaboration is improved and people can better exploit time saved avoiding travels.

A second important aspect to consider to improve collaboration is the ease of contacting a person. We expect that a system for scientific collaboration should be able to keep trace of people movements; depending on where a person is (in his/her office, traveling, etc.) the system might choose the better way to contact him/her (a simple e-mail, or a phone call, or a SMS, etc.). If a person whose role is crucial for a

given task or activity can be contacted quickly, the collaboration is certainly improved and we expect that the overall satisfaction of involved people is improved.

Ease of knowing how the collaboration is going on. This is another important issue. Though the system, people should exchange results about their specific activities; this way, it is possible to understand who is doing what. In fact, knowledge about activities performed by all the partners is crucial to improve the effectiveness of the collaboration, and the system can easily make this knowledge available; for example, in multi-national research projects, periodic reviews are usually the moments when this knowledge is made available, while we expect that through the system this knowledge is made available as soon as a new result is obtained.

This is an important issue also for end users: though the system they can be notified when new documents or data are available, to evaluate if the obtained results are significant or not (for example, a physician might be notified about a new therapy, and may try to understand if the proposed therapy might be adopted or not). This way, end users may become more collaborative with the rest of the team.

Personal satisfaction. Scientific collaboration can be improved only if involved people are satisfied. This fact should be implied by the concept of subtask. In fact, the main idea behind the definition of a sub.task is that the supervisor of the main task delegates responsibility to other research groups involved in the scientific collaboration. At this point, the supervisor of the subtask is free to lead the subtask as he/she prefers, while he/she has to collaborate with the supervisor of the main task to synchronize his/her subtask with the main task. In fact, as far as the overall scientific collaboration is concerned, it is not important how a subtask is conducted, but its results.

Trust about data. For partners which hold data and share them to the other partners, it is important to be sure about the fact that other partners will not make an improper use of shared data. We can expect that a system for decentralized knowledge discovery should provide services for secure data exchange, based on rigid control access policies, should adopt cryptography techniques when data are moved through the Internet, should keep trace of all accesses to the data performed by partners. These are not trivial technical issues; but we think that they should be provided to make effective scientific collaboration through decentralized knowledge discovery.

References

Agrawal, R., T. Imielinski, and A. Swami. Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering, 5(6):914--925, December 1993.

Agrawal R., and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB Conference*, Santiago, Chile, September 1994.

Brugali, D. Mediating the internet. *Annals of Software Engineering*, 13:285--308, 2002.

Coalition, W.M., Information and publications. *http://www.wfmc.org/*.

Gao, L. and X. Wang. Continually evaluating similarity-based pattern queries on a streaming time series. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, June 3-6, 2002, Madison, Wisconsin, USA, 2000.

Imielinski, T. and H. Mannila. A database perspective on knowledge discovery. In *Communications of the ACM*, 39(11):58--64, November 1996.

Jain, A., M. Aparicio, and M.Singh. Agents for process coherence in virtual enterprises. *Communications of the ACM*, 42(3):62--69, March 1999.

Jiawei, H., P. Jian, and Y. Yiwen. Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, May 16-18, 2000, Dallas, Texas, USA, pages 1--12, 2000.

Meo, R., G. Psaila, and S. Ceri. An extension to SQL for mining association rules. *Journal of Data Mining and Knowledge Discovery*, 2(2), 1998.

Psaila, G. Enhancing the kdd process in the relational database mining framework by quantitative evaluation of association rules. In *Knowledge Discovery for Business Information Systems*. Kluwer Academic Publisher, January 2001.

Supporting Scholars' Collaboration in Document Seeking, Retrieval, and Filtering

Sanna Talja
University of Tampere, Finland

Abstract. The study of scholarly communities, their work cultures and information practices, is logically prior to the design of tools for supporting scholars' collaboration in document seeking, retrieval, and filtering (DSR&F). Yet, there is lack of research focusing explicitly on information sharing[1] practices in different fields. This is a challenge for CSCW. This position paper describes some preliminary empirical findings regarding different types of collaborative DSR&F, and discusses their design implications.

Background

The traditional humanistic concept of individuals as the originators of knowledge and of the growth of knowledge as a process initiating from the innovative capabilities of single individuals has tended to dominate both the research on scholars' DSR&F practices and the development of interfaces for document retrieval. Most document retrieval systems (databases, OPACs, digital libraries) interfaces reflect single user stereotypes, and do not adequately support collaboration in the search process (Twidale & Nichols 1998a).

Earlier research has established that the of amount scholars' social ties and memberships in invisible colleges affect information practices so that well-

1 I use information sharing as an umbrella concept covering a wide range of collaboration behaviours from sharing accidentally encountered information to collaborative query formulation, database searching, document filtering, and synthesis.

networked senior scholars receive information of relevant documents through their networks and cite the other members of these colleges (Crane 1972; Allen 1977). However, DSR&F practices are even more fundamentally social. STS studies have shown that scholars' social networks are the *place* where information is sought, interpreted, used, and created. Scientific research is bound up with social interaction. The need to acquire information, to select, distill, and modify ideas, all involve scientists in communication, and "communication is, by definition, a communal activity" (Meadows 1998, p. 49).Although studies on scholars' information practices have firmly established the importance of scholars' social networks for finding relevant literature (Poland 1991) and the general preference of what Selden (2001) calls "social seeking" versus "technical searching," few empirical studies have been conducted on scholars' collaboration in DSR&F.

The study

I gathered the data on scholars' information sharing practices by informal semistructured interviews as part of a larger project, Academic IT-cultures (2000-2002). Four different disciplines, nursing science (a field between natural and social sciences), history, literature/cultural studies, and ecological environmental science (a laboratory science), were chosen as the objects of study in the project. The aim of the comparative ethnography was to develop concepts and hypotheses to be enriched in later studies. The selection aimed at ensuring diversity in the participants' information work methods and types of documents they use. Two humanities fields were chosen because among humanists it was easiest to recruit both participants working alone and researchers involved in research groups. The sample contains 12 nursing scientists, 11 historians, 11 literature and cultural studies scholars, and 10 environmental scientists from two different Finnish universities. Departmental and individual researchers' homepages were also used research data, as they contained valuable information about research groups and research activities. The interviews lasted about 90 minutes, and they were tape-recorded and transcribed in full for analysis.

Empirical findings

Strategic sharing

In the department of nursing science, there was a research project in which one researcher of the team did the initial actual searching on behalf of the whole research group. The leader of the team and this researcher together chose the keywords to be used in searches. These searches were replicated later by research assistants under the project leader's guidance. The information sharing practice adopted in this

research project can be called *strategic* since the progress of research and publications was purposefully designed to function on the basis collaborative DSR&F.

The project leader said that without "designed helps" and well organized "centralized searching" she would never have the time to do actual research. She pointed out that "whatever the area, I want everything that can be found to be taken," and emphasized that centralized scanning makes it possible to cover a larger area, so that "the project files contain everything that can possibly be needed at this moment." Wide-range scanning is necessary in nursing science, because in this multidisciplinary field, relevant documents are scattered across fields such as medicine, education, and social psychology. Two research assistants who also did their own masters thesis as parts of the project were hired to filter and describe the contents of retrieved articles according to a scheme the project leader had designed, "so that I know when I start writing exactly which articles are relevant for that particular piece."

The project team contained 10 researchers and 5 students. Large scientific teams that combine the efforts of several scientists to address a specific problem or problems clearly have the need to share the search process (searches conducted, keywords, and queries), the search product (references obtained), the filtering schemes and filtered results (classified references), and possibly content synthesis schemes and results (systematic literature reviews) among the team members. Although such activities can be performed by using software such as Lotus Notes, a seamless experience requires the integration of these functionalities within IR systems interfaces. The Ariadne project has developed an "interface on interface" solution for collaborative DSR&F (Twidale & Nichols 1998b), another solution is to offer a choice between group searching and individual searching interfaces. Blake and Pratt (2002) have developed a system called METIS for collaborative information synthesis.

Paradigmatic sharing

The research group of digital art and culture, functioning in the department of comparative literature but also attracting members from the university's other departments, started in 1997. The leader of the group and some group members had started doing research on hypertext already in the 1980s. The leader of the group said that in the beginning, the group sought and "knew from a wide sector everything there is." The group members gathered together as a group to identify the classics of the field and the most significant new work. They shared their findings and interpretations concerning important new work and usable older theoretical work not only within the group but also in the group's homepage. They engaged in collaborative seeking, filtering, and interpretation of documents with the aim of connecting and applying existing theories to new topics.

Later, as the interest in digital culture became more widespread in general, and a more "normal" research interest for literature and cultural studies scholars, researchers adopted more specialized viewpoints to digital culture. The group branched off to those studying information technology and those studying audio-visual culture. In the beginning, however, the group members needed each other to collectively develop and establish a shared understanding that information technology, conceptualized as "digital culture," can be a "proper" research interest for scholars in the field of literature. The information sharing practice adopted in this group can be called *paradigmatic*, as the research group commonly strove for a new kind of understanding and definition concerning the subject matter and important research questions in their field, and the most fruitful approaches for studying these questions.

Research groups that form around new topics, subject, approaches, or methods, often cannot use existing keywords to identify relevant documents; they seek to connect older existing documents and theorists to a new or emerging subject area or keyword. In the light of STS theories, all DSR&F takes place within the boundaries of specific schools of thought, paradigms, discourses, and discourse communities. Scholars in multiparadigmatic disciplines especially search for documents with such things in mind, topicality *per se* often being for them only the secondary relevance criterion (Tuominen et al 2003). Standard reference tools (secondary literature databases, digital libraries, thesauri, classification and indexing schemes) do not, however, map the structure of scientific conversations in a particular field (the competing arguments, theories, and research lines) (Agre 1995; Tuominen et al. 2003). Paradigmatic sharing might be supported by tools such as the Claimaker, an experimental system developed for collaborative modeling and visualization of conversations in a particular field (Buckingham Shum et al. 2003).

Directive sharing

In the departments of nursing science and ecological environmental science, graduate and doctoral students were occasionally enrolled in the research projects progressing in the departments. Senior researchers in these departments not only suggested relevant literature to the students, but often also benefited from the searching done by them, and the students benefited from the projects' cumulated document stores. Directive sharing not only took the form of sharing information about documents, but also sharing documents and information about document retrieval techniques. One senior plant researcher told that document retrieval methods "have been taught collegially here, people will tell you that you can find the data you want from there, with that keyword. It has been taught in a mouth-to-mouth fashion." Mentors had relevant articles copied often directly also to their students, or sent their URL addresses by e-mail.

IR interfaces already provide the possibility to email search results to others. Directive sharing could be supported also by other kinds of indirect and

asynchronous collaboration options, like the possibility to share search histories (Komlodi & Soergel 2002) or to request "show me the searches that N.N did" (Twidale & Nichols 1998a). The latter would require prior authorization (the possibility to name those free to view conducted searches).

Social sharing

Respondents all in the studied disciplines often stated in the interviews that sharing information about relevant documents between colleagues is "an extremely good practice that we have in this department." Twidale, Nichols and Paice (1997) call the practice of sharing information about potentially relevant documents between researchers working in different projects and fields "serendipitous altruism." The sharing of accidentally encountered information with others is not strictly goal-oriented, rather, it most resembles the practice of giving and receiving gifts (Erdelez & Rioux 2000). It is a part of building and maintaining collegial relationships, and developing a sense of community where otherwise might only be researchers working alone with their own projects.

In social sharing, information about the contents of documents is less often shared, as scholars may not know exactly how the discussion of a specific document is related to the colleague's topic. Scholars working in the same field but in different research areas cannot necessarily always understand the subtle but essential differences in ways of approaching a particular topic. References coming from respected colleagues or mentors can, however, be more readily judged as relevant (or a relevance may be invented to them) than those found individually by chaining from the bibliographies of seed documents or searching from databases. "Invented relevance" and "relational relevance" are relevance categories rarely identified in relevance and document use literature. In the information retrieval research tradition, DSR&F practices are conventionally assumed to be matters of finding and selecting the topically most relevant documents; yet, especially in human and social sciences, document selection and use are more matters of choosing between different epistemic positions, scientific cultures, and communities a scholar wishes to belong to. This is another reason why humanists and social scientists rarely undertake database searches (Talja & Maula 2003). For them, the Internet is an information seeking environment that fosters and supports accidental encountering of relevant documents (Talja & Maula 2002). Humanist scholars also subscribe to listservs substantially more often than natural scientists, because lists provide access to groups in which epistemic positions are built and discussed (ibid.).

Non-sharing

Non-sharing is combined with research projects so unique that the researchers cannot delegate any part of their information seeking to others, because only they would know when a finding is a finding. In these instances, it is highly unlikely that others could encounter information that would be relevant to these scholars. Such

unique projects are rare, because, as a rule, scholars tend to study that which has already been studied (Bowker 2000, p. 657), or at least use common theoretical and methodological literature. Non-sharers in the study were three historians and literature scholars who did classical humanist research of the "life and works" kind on important historical figures that no one had before written about. Their research was empirical in its character, relying on insight, storytelling and interpretive abilities, or, in a senior historian's words, "normal logic and healthy common sense," more than explicit methods and theories. Their main sources were people who had known the people they were writing about, archive materials, and, in general, documents that could be estimated to contain relevant information, but whose relevance could only be determined by closer scrutiny.

The findings reported here concerned only "naturally occurring" collaborations - those that had evolved from off-line collegial contacts, physical proximity, shared concerns, and apprenticeship relations. Historians and other researchers of historical texts often find themselves alone in their specialty in their departments and universities. Cultural heritage digital libraries as well as research-oriented digital libraries focusing in specific subjects and domains can be designed to support also discussion, collaboration, and matchmaking between users sharing similar interests (Marchionini 1999; Tuominen et al. 2003).

Closing remarks

In their article on collaborative information synthesis, Blake and Pratt (2002) argue that "scientists should make the methods that they use to identify, extract, and analyze information explicit, rigorous, non-biased, and repeatable. Although these traits are the cornerstone of systematic review of biomedical literature, we argue that they are true of good scientists in other disciplines." The preliminary findings of the comparative ethnography on collaborative DSR&F practices described here refute this argument. The findings show that in different disciplines, and depending on the goals of collaboration, and the structure and nature of research teams and projects, different tools are needed to support collaborative DSR&F. Criteria for document selection differ in natural, social, and human sciences. Traditional noun-based indexing languages offer little help for scholars who do not orient to "topically relevant documents" but to "scientific conversations" (Tuominen et al. 2003). The mistake of planning tools to support collaborative DSR&F according to an idealized model of scientific research derived from natural sciences in the same way as in the past happened in the design of reference tools (databases, documentary languages, digital libraries) needs to be avoided.

References

Agre, P.E. (1995) Institutional circuitry: thinking about the forms and uses of information. *Information Technology & Libraries* 14: 225-230.

Allen, T.J. (1977) *Managing the flow of technology: technology transfer and the dissemination of technical information within the R& D organization*. Cambridge, MA: MIT Press, 1977.

Blake, C. & Pratt, W. (2002) Collaborative information synthesis. In Toms, E.G. (ed): *ASIST 2002: Proceedings of the 65th ASIST Annual Meeting*. Medford, NJ:Information Today.

Bowker, G. C. (2000) Biodiversity datadiversity. *Social Studies of Science* 30(5), 643-683.

Crane, D. (1972) *Invisible colleges: diffusion of knowledge in scientific communities*. University of Chicago Press.

Erdelez, S. & Rioux K. (2000) Sharing information encountered for others on the Web. *New Review of Information Behaviour Research* 1(1), 219-233.

Komlodi, A. & Soergel, D. (2002) Attorneys interacting with legal information systems: Tools for mental model building and task integration. In Toms, E.G. (ed): *ASIST 2002: Proceedings of the 65th ASIST Annual Meeting*. Medford, NJ:Information Today.

Marchionini, G. (1999) Augmenting library services: toward the sharium. Paper presented at the internal symposium on digital libraries, ISDL 99, September 28-29, University of Library and Information Science, Tsukuba, Japan. (URL: http://ils.unc.edu/~march/)

Meadows, A.J. (1998) *Communicating research*. San Diego: Academic Press.

Poland, J. (1991) Informal communication among scientists and engineers: a review of the literature. In Steinke, C.A. (ed) Information seeking and communicating behavior of scientists and engineers. New York: Haworth Press.

Selden, L. (2001) Academic information seeking - careers and capital types. *New Review of Information Behaviour Research* 1(2), 195-215.

Talja, S. & Maula, H. (2002) Virtuaalikirjastojen rooli tutkijoiden tiedonhankintakäytännöissä [The role of virtual libraries in researchers' information seeking practices]. *Informaatiotutkimus* 21(2), 35-50.

Talja, S. & Maula, H. (2003): "This field is not based on technical searching:" Disciplinary and intra-disciplinary differences in scholars' use of FinElib's electronic journals and databases. Paper presented at the *Digilib 2003 "Toward a user-centered approach to digital libraries" conference*, 8.-9.9.2003, Espoo, Finland.

Tuominen, K., Talja, S., & Savolainen, R. (2003) Multiperspective digital libraries: The implications of constructionism for the development of digital libraries. *Journal of the American Society for Information Science and Technology* 54(6): 561-569.

Twidale, M. B. & Nichols, D. (1998a) Computer supported co-operative work in information search and retrieval. In M. E. Williams, ed. *Annual Review of ,Information Science and Technology* 33. Medford, NJ: Information Today. 259-319.

Twidale, M.B. & Nichols, D.M. (1998b) Designing interfaces to support collaboration in information retrieval. *Interacting with Computers* 10(2), 177-193.

Twidale, M. B., Nichols, D. M. & Paice, C. D. (1997) Browsing is a collaborative process. *Information Processing and Management* 33(6), 761-783.