COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 07-140 **NSF 07-577** **12/10/07**	**FOR NSF USE ONLY** **NSF PROPOSAL NUMBER**
FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.) **IIS - INFO INTEGRATION & INFORMATICS**	

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
				002257350	

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN) **150532081**	SHOW PREVIOUS AWARD NO. IF THIS IS □ A RENEWAL □ AN ACCOMPLISHMENT-BASED RENEWAL	IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES □ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE **Syracuse University**	ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE **Syracuse University** **Office of Sponsored Programs** **Syracuse, NY. 132441200**
AWARDEE ORGANIZATION CODE (IF KNOWN) **0028829000**	
NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE	ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE
PERFORMING ORGANIZATION CODE (IF KNOWN)	

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions) □ SMALL BUSINESS □ FOR-PROFIT ORGANIZATION □ MINORITY BUSINESS □ WOMAN-OWNED BUSINESS □ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **III-CXT - Small: Semi-automated coding of qualitative data to study group maintenance in self-organizing distributed teams**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **445,780**	**24** months	**07/01/08**	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

□ BEGINNING INVESTIGATOR (GPG I.G.2)

□ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C)

□ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.D, II.C.1.d)

□ HISTORIC PLACES (GPG II.C.2.j)

□ SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.D.1)

□ VERTEBRATE ANIMALS (GPG II.D.5) IACUC App. Date ________
PHS Animal Welfare Assurance Number ________

☒ HUMAN SUBJECTS (GPG II.D.6) Human Subjects Assurance Number ________
Exemption Subsection ________ or IRB App. Date **01/01/06**

□ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
____

□ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

PI/PD DEPARTMENT **School of Information Studies**	PI/PD POSTAL ADDRESS **Syracuse University Information Studies** **348 Hinds Hall** **Syracuse, NY 132444100** **United States**
PI/PD FAX NUMBER **315-443-5806**	

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME **Kevin Crowston**	**PhD**	**1991**	**315-443-1676**	**crowston@syr.edu**
CO-PI/PD **Ozgur Yilmazel**	**PhD**	**2006**	**315-443-2807**	**oyilmaz@syr.edu**
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

Project Summary: III-CXT - Small: Semi-automated coding of qualitative data to study group maintenance in self-organizing distributed teams

This study explores the application of Natural Language Processing (NLP) and Machine Learning (ML) tools to the context domain of organizational behaviour, more specifically to a study of group maintenance in a novel setting. The proposal involves information scientists working collaboratively with domain scientists with goal of developing an innovative NLP and ML-based research tool to support qualitative social science research, specifically content analysis. Content analysis is a qualitative research technique for finding evidence of concepts of interest using text as raw data rather than numbers [75]. The process of identifying and labelling significant features in text is referred to as "coding" and the result of such an analysis is a text annotated with codes for the concepts exhibited [72]. The problem of coding qualitative data is conceptualized as an Information Extraction (IE) problem. However, rather than seeking to automate the process, the system will employ the technologies in a supporting role, keeping the human coder in the loop. Specifically, it will apply an active learning process, using a few hand-coded examples to create an initial model that is evolved through interaction with the user. The project thus advances the domain by allowing qualitative researchers to obtain the benefits of cyber-infrastructure in leveraging their research capabilities. To validate the utility of the tool and further advance the domain, the system will be applied to the study of group maintenance behaviour in cyber-infrastructure-supported distributed groups, specifically free/libre open source software development teams.

Expected intellectual merit

The intellectual merit of the proposed research is three-fold. First, the innovative information science contribution of the proposal is the integration of information extraction and active learning in an interactive system to reduce the required amount of hand-annotated training data for the information extraction system, which will make practical the use of a system for coding qualitative data in various domains. A validation study will apply the tool to a diverse set of codes, providing evidence of the generality and limits of the approach. Second, the project addresses a fundamental methodological problem in the broad domain of qualitative research, namely dealing with large quantities of unstructured qualitative data, by applying innovative information extraction technologies. Finally, a second domain science contribution of the study is to address a fundamental problem in the application domain of organizational behaviour, namely group maintenance in a novel setting, namely distributed groups working together using cyber-infrastructure. The study will contribute by advancing our understanding of the effects of interpersonal relationships on the functioning, effectiveness and innovation of groups who rely on innovative applications of computer-mediated communications (CMC).

Expected broader impacts

The project has numerous broader impacts. In addition to the expected intellectual contributions described above, the proposed research will benefit society by providing a component of useful infrastructure for qualitative science research, thus contributing to the infrastructure of science. In particular, the toll will be integrated with cyber-infrastructure currently being developed by one of the PIs with other NSF support. A second goal is to provide generalizable knowledge to improve the effectiveness of distributed groups, a further benefit to society. Such groups are an increasingly important approach to needs such as software development, scientific research and policy development. Distributed work is potentially transformative for organizations and society, but the separation between members of distributed groups creates difficulties in building social relations, which may ultimately result in a failure of the group to be effective. For the potential of distributed groups to be realized, research is needed on how to make them engaging and motivating to members. Understanding the role of group maintenance in these settings and the relation to group performance will help us develop guidelines to improve performance and foster innovation.

Keywords: qualitative content analysis; group maintenance; information extraction; active learning

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.C.

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

III-CXT - Small: Semi-automated coding of qualitative data to study group maintenance in self-organizing distributed teams

We propose a study that explores the application of information science, specifically Natural Language Processing (NLP) and Machine Learning (ML), to the context domain of organizational behaviour, more specifically, a study of group maintenance in a novel setting. The proposal involves information scientists working collaboratively with domain scientists with goal of developing an innovative NLP and ML-based research tool to support qualitative social science research, thus advancing the domain by allowing qualitative researchers to obtain the benefits of cyber-infrastructure in leveraging their research capabilities. The innovative information science contribution of this proposal is the integration of information extraction and active learning in a tool to support a commonly applied qualitative data analysis approach, content analysis, an example of extraction of structured information from unstructured sources. To validate the utility of this tool and advance the domain, we will apply the tool to the study of group maintenance behaviour in cyber-infrastructure-supported distributed groups.

The proposed project has two overlapping phases, each lasting roughly 12 months and each including work in the domain, in information science and in the application of information science to the domain. In the first phase (year 1), for the domain, we will use computer-assisted qualitative data analysis software (CAQDAS) to examine project data to determine what kinds of group maintenance seem most important in our groups and to elucidate the antecedents and outcomes of group maintenance behaviour in order to propose specific hypotheses for further study. In parallel, for the information science contribution, we will examine how the concepts of interest are linguistically realized in text to determine feasible candidates for identification using NLP techniques. Based on this analysis, we will develop an information extraction system and implement active learning algorithms that can partially automate qualitative data analysis. In the second phase (year 2), we will integrate the NLP and ML algorithms into a working prototype research support system, which we view as a component of cyber-infrastructure for the domain. We will then use the system, seeding it with the group maintenance data manually coded in year 1 and iteratively training it to analyze group maintenance in a large number of projects, thus assessing the relation of group maintenance to group effectiveness in this setting. The experience of using our newly developed system on a large body of data and with a diverse set of codes will provide: 1) validation of the utility of the system and the approach, 2) information about which kinds of concepts are more or less amenable to the proposed approach and 3) suggested enhancements to the system.

The intellectual merit of the proposed research is three-fold. First, the innovative information science contribution of the proposal is the integration of information extraction and active learning in an interactive system to reduce the required amount of hand-annotated training data needed for the information extraction system, which will make practical the use of a system for coding qualitative data in various domains. A validation study will apply the tool to a diverse set of codes, providing evidence of the generality and limits of the approach. Second, the project addresses a fundamental methodological problem in the broad domain of qualitative research, namely dealing with large quantities of unstructured qualitative data, by applying innovative information extraction technologies. Finally, a second domain science contribution of the study is to address a fundamental problem in the application domain of organizational behaviour, namely group maintenance in a novel setting, namely distributed groups working together using cyber-infrastructure. This study will advance our understanding of the effects of interpersonal relationships on the functioning, effectiveness and innovation of groups who rely on innovative applications of computer-mediated communications (CMC). The distance (physical, organizational and social) between distributed members and the limited opportunities for interaction provided by cyber-infrastructure suggest that many of the traditional tactics of group maintenance will be difficult to apply, but “the social glue of good relations among participants” is still critical [7].

The remainder of this proposal is organized into four sections. In section 1, we discuss the process of qualitative data analysis and the increasingly pressing problems faced by qualitative researchers interested in cyber-infrastructure. We then discuss the promise offered by NLP and how NLP tools might be used to support qualitative data analysis. In section 2, we discuss the specific research setting that will be used as an application domain for the proposed study and present the study design, with details of the data collection and analysis plans. In section 3, we present the project management plan and requested resources.

We conclude in section 4 by describing the intellectual merits and expected broader impacts of our study and by reviewing results of prior NSF support.

1. The problem of qualitative research and a possible solution

Social science researchers often study interpersonal communication in order to understand the practices of the populations in which they are interested. Such data are typically textual in nature and are therefore not directly amenable to quantitative analysis. To ease the analytical process, many researchers use Computer Assisted Qualitative Data Analysis Software (CAQDAS) tools, but these tools do not yet offer all of the capabilities researchers need, and have not reached the level of sophistication and automation that quantitative tools have reached [8, 11, 74, 102]. To put it bluntly, qualitative analysis does not scale—rather, it is limited by the capabilities of individual researchers. Important research questions in the qualitative social sciences may rely on insufficient sample sizes because of the limitations of the tools and their need for intensive human effort, or worse yet, they may fail to be addressed in studies at all. Moreover, dissemination of findings is delayed due to time and effort it takes to analyze data.

The problem described above is only getting worse. Organizations of all types and sizes are increasingly using various forms of technology-supported collaboration [13, 47, 53] and the distributed collaborative practices of these groups produce an enormous amount of digital data, such as e-mail listserv archives, instant messaging logs and weblogs. These digital data sources have the potential to augment traditional sources such as interview transcripts and participant observation notes. If fully exploited, this digital data could make a rich contribution to qualitative social science research addressing individual and group behaviour in technology-supported groups, and by extension, group and organizational behaviour more generally. In fact, if the social sciences are to move from ‘Little Science’ to ‘Big Science’ the way the physical sciences have [37], qualitative social science researchers need to exploit a fuller range and volume of data collection and analysis that improved tools would make available to them. However, the sheer volume of the data poses significant challenges. It is not uncommon for mailing lists to have dozens of emails per day, and the concepts of interest are often only indirectly reflected in the behaviours.

Our research proposal is based on the belief that Natural Language Processing (NLP) and Machine Learning (ML) techniques can provide advanced analytic capabilities to assist qualitative social science researchers in analyzing large volumes of data. Such innovative tools offer the promising of extending the depth, breadth, and efficiency of qualitative data analysis and optimally utilizing researchers’ intuitive and analytical skills by leveraging the large-scale processing capabilities of computers to deal with vast repositories in consistent, reproducible ways. Of course, it is unrealistic to expect such tools to automate analysis—instead, just as for quantitative data, tools must be developed that support the researcher and make large volumes of data more accessible. If successful, these sophisticated NLP tools will advance the work of qualitative social science researchers by extending the capabilities of current CAQDAS tools and enabling researchers to explore massive amounts of data in more complex ways.

In the remainder of this section, we first review the process of content analysis, which we illustrate with an extended example drawn from our prior NSF-supported research. We then discuss how NLP tools might be applied in this domain, again providing an illustration of some pilot research in our current project. We finally discuss the approach proposed for the current proposal, namely using active learning to iteratively induce rules for coding and describe the proposed implementation.

Content analysis as a qualitative analysis approach

In this proposal, we will harness NLP and ML techniques to support the process of qualitative research, specifically, to support the process of content analysis. Content analysis is a qualitative research technique for finding evidence of concepts of interest using text as raw data rather than numbers [75]. Content analysis of computer-mediated communication in particular has been an active area of research [6, 55]. It is commonly assumed that qualitative work must be interpretivist (i.e., concerned with describing individuals’ understandings of their social worlds), but in fact qualitative research can adopt any research perspective: positivist, interpretivist or critical [75]. In particular, for the purposes of the proposed study, we assume that the nature of group processes are accurately reflected in the texts group members produce, making our approach essentially positivist. This analysis approach has advantages in that it does not require the active participation of the individuals being studied, which can be difficult to obtain if they are busy or no longer available, nor does it rely on participants’ possibly fallible recollections or impressions of the process. On the other hand, the understanding we develop by analyzing the process from an external (or “etic”) perspective may not be the same as the understanding participants have themselves (an “emic” perspective). By contrast, it is typical for interpretivist or critical analysis to augment observational data with interviews to develop “rich descriptions” of the setting. Such an analysis would aim at uncovering hidden meanings in the texts rather than evidence of pre-specified concepts, something that would be much harder to automate, and so beyond the scope of the current proposal.

The process of identifying and labelling significant features in text is referred to as “coding” and the result of such an analysis is a text annotated with codes for the concepts exhibited [72]. A codebook documents the coding process by describing the characteristics of the text that count as evidence for each concept of interest. A codebook might also include definitions or references for the concepts represented and positive and negative examples of text that is evidence for a code, although it has to be admitted that much of the knowledge that guides coding is held tacitly by the coders. A key concern in developing a codebook is its reliability, i.e., the degree to which different coders working with the same text identify the same set of codes, as measured by the degree of inter-rater agreement. If coders do not agree, then it is typical to have them discuss the coding until they reach a higher level of shared understanding of the code and to update the codebook accordingly. The coded text can then be subject to further analysis, such as examination of the relationship between codes or quantitative analysis of their co-occurrence. Content analysis can be deductive or inductive or most often, a mix. A deductive approach is based on a theoretical framework that identifies concepts of interest for the codebook. Such an approach would be appropriate when the goal of the analysis is to test the theory. A pure inductive approach starts with a research problem and data, and induces relevant concepts from them, setting aside any preexisting concepts [41]. Such an approach is appropriate when the goal is developing novel theory for some unexplored setting. A mixed mode analysis, probably the most common approach, starts with relevant concepts from theory, but allows these to evolve through the analysis based on experiences with data.

Results from prior funding: Content analysis of decision process in FLOSS teams

To make the discussion of the qualitative content analysis coding process more concrete, we present a specific example drawn from a study currently under way by one of the PIs, supported by NSF Grant HSD 05–27457, *Investigating the Dynamics of Free/Libre Open Source Software Development Teams* (with R. Heckman and E. Liddy). The overall goal of that project is to examine the evolution of effective work practices in a particular kind of distributed team, namely teams of software developers working on free/libre open source software [18]. As part of this study, we focused on the decision-making processes in teams [50, 51] (other aspects of the study are reviewed below under results from prior funding). Literature suggested that the process by which groups reached decisions would have important consequences for team effectiveness but that the distributed and voluntary nature of these teams would make effective processes difficult to achieve.

To find evidence describing these processes and their link to effectiveness, content analysis was applied at multiple levels. First, transcripts of team email discussions were read to find examples of the team facing a situation that required a decision (a “decision trigger”) and for announcements of decisions that had been made (a “decision announcement”). We created codebooks for triggers and announcements, based initially on the literature, and evolved them as different kinds of triggers and announcements were identified. For example, a trigger might be a bug report that requires a decision about a code change or a proposal to add a new developer and the corresponding decision the acceptance or a patch or of the developer. Once triggers and announcement were identified, we coded them on various dimensions, e.g., what kind of trigger or the status of the person sending the message (a project administrator, a developer or user). The trigger and announcement were used to identify the stream of messages that included the decision process. Individual messages were then coded for the stage of the process involved (problem exploration, solution formulation, solution evaluation, selection, a framework drawn from the literature) using a further codebook. Episodes were then categorized based on the nature of the process (e.g., linear or iterative, complete or partial). Finally, participation in and the nature of decision-making processes were related to overall team effectiveness. An example finding from this work was that the less effective teams had decision-making processes characterized by lower levels of user participation.

Because the process of coding involves careful reading of texts to find instances of the phenomena of interest, it is extremely labour-intensive. The study described above produced only 360 coded decision episodes (6 projects x 3 time periods per project x 20 episodes per period) but coding required nearly 1 person year of effort (2 half-time GAs for a year with additional support from other researchers for conceptual development). Part of this time was spent developing and refining the codebook, but much of it was spent reading and rereading messages, coding examples of the phenomenon of interest. The resulting dataset is suggestive, but the small number of projects included does not support firm conclusions about the hypotheses of interest. On the other hand, to increase the number of projects to 60 (a sufficient sample size for statistical analysis with the power needed to draw conclusions) would require a prohibitive amount of labour, even with the developed codebook. This situation—a high input of labour for a small payoff of data, and thus limits on the kind of research question that can be addressed—is the domain problem we address in this proposal.

Automated supported for qualitative content analysis coding

To support qualitative analysis and address the problem identified above, we propose applying Natural Language Process (NLP) and Machine Learning (ML) technologies. We conceptualize the problem of coding qualitative data as comparable to an Information Extraction (IE) problem. IE is a subfield of NLP whose function is to extract or annotate desired parts of unstructured text, thus extracting structured information from unstructured information. Rather than seeking to automate the coding process, we will employ the technologies in a supporting role, keeping a human coder in the loop. The innovative information science contribution of this proposal is the integration of information extraction and active learning in a tool to support content analysis. Specifically, we will implement a Web-based information extraction service for researchers in which a researcher can upload their unstructured data, interactively define the information extraction schema of interest and then use the system to identify additional examples of the concepts. A human coder would review the coded data during the coding process and before being it is used for further analysis.

Results from prior funding: Initial experiments with qualitative content analysis coding using NLP

Information Extraction systems are of two types, rule-based and statistical-learning-based. A rule-based Information Extraction system relies on an expert to write rules to capture the intended schema. These rules are usually topic and domain dependent, which makes using the system for a different schema hard. As part of our work on the HSD grant mentioned above, we have experimented with rule-based NLP for qualitative data coding. The rules were developed iteratively by a trained NLP analysis working with the human coders. Rules were coded using part-of-speech information, word order, semantic class, and domain-based world knowledge to extract the decision triggers and announcements similar to those manually coded. The development optimized for coverage of episodes (i.e., recall, extracting the majority of the decision announcements in the data) rather than precision (ensuring that the extracted decision announcements were all correct), under the assumption that coders can more easily discard incorrectly coded segments than they can search the entire email logs to find the decision triggers and decision announcements not identified by the system.

To evaluate performance, the system's output was compared to the manual coder's output, which was assumed to be correct. After several rounds of iteration, the goal of achieving good coverage (recall) with acceptable precision was reached. Tables 1 and 2 report performance when using the final rule set with messages from the developer mailing list of 3 Internet messaging client projects (Fire, Gaim and aMSN). We report the traditional information extraction metrics of recall and precision, where recall measures the proportion of manually coded decision announcement statements that were correctly extracted by the system and precision measures the proportion of extracted decision announcement sentences that matched those manually coded. In addition, we report utility [9], which assesses the benefit or usefulness of the tagged instances to the researcher. The metric is difficult to quantify but important to consider when trying to establish whether adding NLP processing to the research process is actually helpful. For this study, utility was defined as the proportion of decision trigger and decision announcement statements that were determined to be useful to the researchers in their understanding of the decision process, as assessed by the analyst. For example, in the data, various sentences, seen out of context, are perfectly good examples of a decision triggers or announcements, e.g. "I just committed the fix". However, the manual codebook being used stated that for decision announcements, the final decision announcement is the one to be coded, not earlier messages, on the grounds that those earlier messages do not represent the final decision.

The utility measure though includes the full range of such statements. Note that the utility achieved with the rules is in the same range as the target 80% agreement often used as a rule of thumb for acceptable inter-rater reliability using human coders, suggesting that automated coding can be useful. Our future plans in the HSD grant are to use the rules to code a large number of projects in order to provide more conclusive results to the study described above. Even with low precision, by identifying the 1% of sentences likely to contain a code, the pilot system could potentially improve coding productivity by two orders of magnitude, allowing us to analyze hundreds of projects.

Developing the pilot system exposed some mismatches between human and machine coding that will be addressed in the proposed research. A specific example of a difference in approach is in the choice of unit of coding. In manual coding, it is common to use the thematic unit as the unit of coding. For example, several sentences might be coded together as a decision announcement, if related. However, for automated coding, the unit needs to be unambiguously identifiable, e.g., the sentence or message. This example illustrates the way the two approaches need to inform each other to work together.

To summarize, our rule-based NLP coding experiments suggest the promise of the general approach to be followed in this proposal. However, they also reveal a major drawback, namely, the skill and effort necessary to develop the rules. Adopting this approach would replace the current qualitative analyst bottleneck with an even more serious NLP analyst bottleneck. We anticipate using rule-based coding for some fixed aspects of messages (e.g., coding message time, subject, sender, receiver, which are less ambiguous and useful for most analyses) but the approach will not scale for coding more varied theoretical constructs used only in particular studies.

The proposed approach: Active learning of coding rules

Fortunately, there is an alternative to manual rule development that we will explore in our study, which is to apply a statistical-learning-based approach to develop rules. In the proposed research, we will experiment with techniques including Hidden Markov Models (HMM), Maximum Entropy [12] or Support Vector Machines [95] for token-classification-based information extraction. We choose to use statistical learning rather than rule learning techniques because of their robustness over different document structures. Token classification in its simple form is a classification task that assigns labels to individual tokens in a document. When modeled as a token classification task, the information extraction task is to assign a semantic label to a given token, identifying if that token is a valid start of an instance of the desired information, the end of an instance or a continuation within an instance. The learning process requires a set of correctly coded or annotated data to represent ground truth. The correct data are divided into two sets, a training set used to build a model to do extractions and a testing set used to evaluate the resulting model. The annotations in the training set are represented in a vector form of attributes for each instance. The attributes can be individual words, context, part-of-speech and any other information we can gather from text.

Although machine learning from instances is easier than developing specific rules for an extraction

Decision Triggers	FIRE	Gaim	aMSN
# manually coded decision trigger sentences	133	98	92
# decision trigger sentences extracted	122	109	112
Precision	71 %	64 %	59 %
Recall	65 %	63 %	72 %
Utility	88 %	83 %	85 %

Table 1. Automatic coding of Decision Triggers – Performance at sentence level

Decision Announcements	FIRE	Gaim	aMSN
# manually coded decision announcement sentences	109	98	113
# decision announcement sentences extracted	136	132	129
Precision	68 %	58 %	62 %
Recall	85 %	79 %	71 %
Utility	84 %	80 %	79 %

Table 2. Automatic coding of Decision Announcements – Performance at sentence level

system, it still requires considerable manual coding of data before learning can start [36]. The problem of getting large amounts of coded data for each domain is the main bottleneck of learning-based IE systems. To reduce this bottleneck, we will apply Active Learning, which is an expansion of the supervised learning process to reduce the required size of the training set [36, 91]. In the active learning process, a few hand-coded examples are used to create a model and the model is then run over the test documents. The system can then ask the user to annotate more data, e.g., the instances that it is least certain about. The user can also choose to correct other annotations or create new ones. Newly annotated data are fed back into the training set and a new model is created. As a result of this focused coding, the system performance improves quickly with fewer training examples. This active learning can continue until the user is satisfied with the output or a certain predefined performance measure is reached. The overall architecture of the approach is shown in Figure 1.

Active learning can use many different techniques to select specific documents or instances for manual annotation [35]. Finn and Kushmerick [35] in their study suggested six document selection strategies suitable for information extraction from text and text classification. Three out of the six suggested are applicable to generic information extraction problems that we are interested in, namely Compare, ExtractCompare and Bag. Compare, as the name suggests, chooses a document to code that is least similar to the training set. ExtractCompare applies the learned extractions over the document and compares the results of the extraction to the training set and selects the document whose extractions are least similar to the training set. Bagging requires dividing the training set into different parts and building models by using these parts. Information extraction is done over documents by using these distinct models, and documents with the least agreement from different extraction systems are selected for further annotation. In our system we will implement these selection strategies and evaluate their effectiveness in reducing the number of annotated examples needed.

2. Domain study: Group maintenance in self-organizing distributed teams

As a domain application to validate the utility of the proposed system, we propose a study of the role of group maintenance behaviours for effectiveness of distributed teams. This work will provide a range of qualitative analysis challenges to demonstrate the utility of the proposed tool as well as complementing the work already underway by one of the PIs examining the dynamics of effective work practices in distributed teams, briefly described above.

Theoretical foundation

We define group maintenance behaviour as discretionary, pro-social, relation-building behaviour that

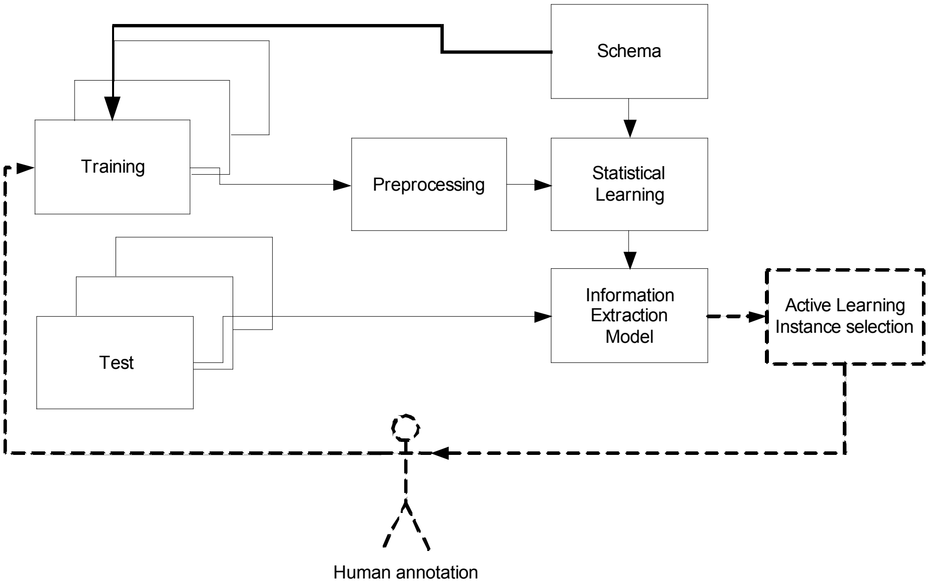


Figure 1. Active learning for information extraction.

is not explicitly task oriented. Though different research streams have used different labels, researchers have commonly differentiated between two broadly defined types of group behaviour: *task-oriented* behaviour and *relational* or *group maintenance* behaviour. Such behaviour is closely related to an array of prosocial behaviors that have been identified by organizational theorists in various contexts: consideration, expressive behaviour, or relational behaviour in leadership research [57, 106, 107]; social presence in community of inquiry literature [38, 90]; social-emotional behaviour, face work, or social presence in CMC research [43, 73, 78]; and organizational citizenship behaviour (OCB), relation-oriented behaviour, supportiveness, conflict management in organizational research [40, 52, 82]. Group maintenance behaviour enables group members to more easily trust and cooperate with one another, based on the expectation of the future cooperation of others [88], what game theorists call the "shadow of the future" [4]. Voluntary groups, whether part of businesses, societal communities or research communities, will not last long if members are dissatisfied and ineffective collaborators. Groups that endure develop a social environment that is conducive to accomplishing group tasks, and to the social needs of individual members. This social environment includes open communication among the group members, support of the group members' needs, an effective conflict-resolution process and commitment by the group to minimize process losses [i.e, group synergy, as defined by 46]. Developing a supportive environment is particularly hard in distributed groups, since members have few opportunities to meet and work together face-to-face.

In the remainder of this section, we develop the conceptual framework for the domain study, building on and adding to existing literature drawn from multiple disciplines. While there is no comprehensive theory of group maintenance behaviour, researchers have identified an array of discretionary, pro-social behaviors that contribute to the creation of an environment that supports a work group's task-related activities. We first discuss research in group leadership and organizational citizenship behaviour that helps us to understand the general nature of group maintenance behaviour. We then turn to research that more specifically addresses group maintenance behaviour that might be expected to occur via cyber-infrastructure, as in distributed groups. Finally, we consider group performance literature to address the link from group maintenance to performance.

Group leadership theory. The group leadership literature provides a first perspective on the nature of group maintenance behaviors. Most group leadership studies have adopted a two-factor theory of leadership derived from Bales [5] research on small team interaction, which distinguishes between task- and relationship-oriented leadership behaviour. Task-oriented behaviors are those that move the team forward in the accomplishment of its task, such as "planning and scheduling work, coordinating subordinate activities, and providing necessary supplies, equipment, and technical assistance" [106]. Relationship-oriented behaviors, on the other hand, are those that allow the team to maintain a positive psycho-social dynamic, such as "showing trust and confidence, acting friendly and considerate, trying to understand subordinate problems, helping to develop subordinates and further their careers, keeping subordinates informed, showing appreciation for subordinates' ideas and providing recognition for subordinates' accomplishments" [106], which we consider as group maintenance. In research on self-organizing teams, group leadership has often been described as shared [85] or distributed [45]. Thus, in such groups, we expect that group maintenance leadership behaviors will be performed by a number of group members with a variety of targets, not just leaders to subordinates.

Organizational citizenship behaviour. An additional source of ideas about the nature of group maintenance behaviors is the work on organizational citizenship behaviour (OCB), which has been defined as "individual behaviour that is discretionary, not directly or explicitly recognized by the formal reward system, and in the aggregate promotes the efficient and effective functioning of the organization" [82]. Several dimensions of OCB have been identified, including *helping* (behaviour in which the immediate beneficiary is a specific individual person), *compliance* (general adherence to the spirit of the rules or norms that define a cooperative system), *sportsmanship* (putting up with minor grievances and inconveniences without complaining), *civic virtue* (responsible, constructive involvement in governance processes) and *courtesy* (avoiding practices that make other people's work harder) [44, 63, 81, 94, 104]. This research suggests that OCB is closely related to positive attitudes such as job satisfaction. Theorists have also proposed that dispositional traits (i.e., personality) predict OCB, but the bulk of the empirical research on this issue does not support this relationship [82].

In summary, research on group leadership theory and organizational citizenship behaviour suggest that group maintenance behaviors may be widely performed and has identified a variety of behaviors that may contribute to the development and preservation of a positive environment. However, in distributed groups, the opportunities for group maintenance behaviour are limited because interactions are predominantly mediated by CMC. Heckman and Annabi [49] suggest that the lack of informal, face-to-face communication presents challenges for collaboration and learning in distributed groups, since not all of these behaviors translate to the new environment. In the remainder of this section, therefore, we turn to research that has attempted to identify group maintenance behaviour carried out via CMC, in order to address our first research question. We first briefly review research on *virtual teams* before turning to research on computer-mediated asynchronous discourse, specifically, *community of inquiry* and *politeness theory*.

Research on virtual teams. Martins, Gilson & Maynard [71] recently surveyed the growing body of research on virtual teams (VT), which they defined as "teams whose members use technology to varying degrees in working across locational, temporal, and relational boundaries to accomplish an interdependent task" (p. 808). They found that the "majority of VT research pertaining to interpersonal processes… focused on conflict, uninhibited behaviour…, informality of communication among group members, interpersonal trust, and group cohesiveness" (p. 814). Trust (one of the outcomes of group maintenance behaviour) in particular has a rich literature. For example, Jarvenpaa and Leidner [62] identified what they called "swift trust" that formed in temporary distributed groups. However, Martins et al. note that much of this work has been done in a lab setting with student groups [71, p. 822], which is consistent with a focus on temporary teams. Such research needs to be followed up with studies of longer-standing functioning distributed groups, in particular because experience working together may be a key factor in developing relationships. They further note that "interpersonal processes represent an area in which major gaps exist in the literature on VTs." (p. 821), suggesting a need to consider the specific behaviors that help build relationships and which are feasible in CMC-mediated interaction.

Research in computer-mediated communications. To help identify behaviors that might support group maintenance in a CMC-supported group, we turn now to work that has examined CMC interaction in more detail. The notion of a *community of inquiry* has its antecedents in the work of the American pragmatists in general, and especially John Dewey [34, 79], and the term achieved wider usage through Matthew Lipman's Philosophy for Children movement [70]. A community of inquiry is characterized by trust and an open, critical, collaborative search for meaning and truth. Anderson, Archer, Garrison and Rourke [2, 38, 39, 90] have developed and validated a content analysis scheme to evaluate the learning process of individuals using asynchronous technology to collaborate in a community of inquiry. Building on social interdependence, critical thinking, and constructivist learning theories [38, 48, 56, 76, 80, 103] they presented a model that integrates cognitive presence, social presence, and teaching presence. Their framework identifies the intellectual content of messages (*cognitive presence*), the instructional role (*teaching presence*), and interaction among members (*social presence*). Aviv [3] also developed a framework to analyze the content of messages and the nature of interactions. His framework identifies three processes in asynchronous learning network discussions: *social process*, *response process* and *reasoning process*. These frameworks provide a useful starting point for the identification of group maintenance behaviour in asynchronous communication.

Politeness theory. A second stream of research that provides useful insights into group maintenance behaviour embedded in speech is *politeness theory*. Politeness theory considers the role of *face*, the positive self-image claimed and presented to the social world by each individual [42]. The theory posits that face-threatening acts (FTA) are an inherent and unavoidable aspect of any human interaction using language. Politeness in language represents an effort to support and preserve the self-esteem, or face, of others, to minimize the impact of face-threatening acts. Politeness tactics can be either specifically positive or negative [10]. Negative tactics attempt to avoid negative face by demonstrating distance and circumspection to the other [73]. Positive tactics indicate an appreciation of the other's wants in general [73]. Positive politeness tactics help group members to bond and to locate common ground whereas negative politeness tactics prevent group members from coming too close or intruding by keeping appropriate distance. Based on the work of Brown and Levinson [10], Morand and Ocker [73] developed a set of indicators of positive and negative politeness tactics for use in analyzing CMC transcripts.

We plan to build explicitly on both the community of inquiry and politeness theory frameworks because prior research in these areas has identified linguistic markers that enhance the social dimension of collaboration, suggesting that these concepts will be good candidates for NLP analysis. Table 3 presents a preliminary set of group maintenance indicators identified for research on community of inquiry and politeness theory that we expect to see expressed in cyber-infrastructure-supported communications. The table includes a range of indicators that make different tradeoffs between reliability and validity, i.e., some are very explicit and thus easy to recognize automatically but perhaps only indirect indications of group maintenance, and vice versa. The indicators in the table will thus provide a good test of the range of constructs for which automated identification is feasible and useful.

Group effectiveness. To motivate the study of group maintenance, we will evaluate the relationship between it and group effectiveness. The group performance literature suggests the importance of group maintenance and its relation to other group processes. Research has empirically linked group maintenance behaviour in face-to-face groups with several indicators of positive group or organizational performance. OCB has also been associated positively with performance quantity and quality, financial efficiency, and good customer service [82]. For example, organizational citizenship behaviour has been associated positively with performance quantity, performance quality, financial efficiency, customer service, and attitudes such as job satisfaction [82]. Thus we find a large body of research that associates discretionary pro-social organizational behaviour with desirable group outcomes and characteristics. Because the majority of this research has been cross-sectional and correlational, theorists have been careful to point out that we cannot say with certainty whether variables such as job satisfaction are antecedents of these behaviors, outcomes of these behaviors, or, together with these behaviors, caused by a third variable. Nevertheless, evidence for a relationship between this form of group maintenance behaviour and positive group outcomes continues to grow.

To measure team effectiveness, we will consider outcomes along the three dimensions suggested by Hackman [46]: task performance, as measured by evaluations by recipients of the output (which may include the team members themselves), individual group member satisfaction and continued group performance. For the FLOSS setting, Crowston et al. [23] have developed a set of indicators of effectiveness, including releases and bug fixes as measures of task performance, individual developer satisfaction with the project, and number of developers involved and level of activity as indicators of continued group performance. We anticipate that the effects of group maintenance behaviour will be more visible in certain of these outcomes, e.g., we expect it to have a large impact on the group's ability to retain members.

Table 3. Initial constructs and indicators of group maintenance behaviour**.**

Category	**Indicators**	**Definition**
Emotional expression [38, 90]	Expressions of emotion using emoticons.	Expressions of emotion using emoticons
	Expressions of emotion using conspicuous capitalization.	Expressions of emotion using conspicuous capitalization
	Expressions of emotion using repetitious punctuation	Expressions of emotion using repetitious punctuation
	Explicit expression of emotion	Direct or explicit expression of emotion using emotional words
	Use of humor	Teasing, cajoling, irony, understatements, sarcasm
Positive face protection	Communication commonalities	-Spelling out phonological slurring -Using colloquialisms or slang -Use of group-specific jargon, language, or metaphors
	Cohesion/inclusion	-Use of vocatives (referring to participants by name, or addressing part of a message to an individual) -Use of inclusive pronouns (incorporating writer and recipient[s]) -Use of phatics (personal greetings and closures, including communication for purely social reasons)

	Personal connection (humanizing the exchange)	-Raising/presupposing commonalities -Complimenting others or message content -Expressing agreement with others -Expressing a reciprocal exchange -Expressing empathy and understanding -Apologizing -Encouraging others, calling on others to participate -Repetition to indicate acceptance or idea sharing -Self-disclosure
	Minimizing face threat	-Use of disclaimers prior to an FTA -Stating an FTA as a general rule to minimize impact -Explaining the reasons behind an action -Use of hesitation in disagreement (e.g., “well…”)
Negative face protection	Indirection	-Inquiring into hearer’s ability/willingness to comply -Use of hedges (words/phrases to diminish force of act) -Use of the subjunctive in requesting assistance -Use of phrases to minimize the imposition -Self-depreciation -Avoiding request despite obvious need to make one
	Formalities	-Use of honorifics (Mr., Mrs., Dr., etc) -Using formal verbiage -Impersonalization (avoiding use of I or you) -Use of past tense to create distance
Organizational citizenship behaviors	Helping	Behaviour involving voluntarily helping others with a work problem. The immediate beneficiary is a specific individual person [83] - Helps others who have been absent - Helps others who have heavy work loads - Helps orient new people even though it is not required - Willingly helps others with work related problems - Always ready to lend a helping hand [84]
	Courtesy (viewed as helping [83])	Subsumes all of those foresightful gestures that help someone else prevent a problem; avoiding practices that make other people’s work harder - Takes steps to minimize problems with other workers - Mindful of how behaviour affects other people's jobs - Does not abuse the rights of others - Tries to avoid creating problems for coworkers - Considers the impact of actions on coworkers [84]
	Peacemaking (viewed as helping [83])	Actions that help to prevent, resolve, or mitigate unconstructive interpersonal conflict - Acts as a "peacemaker" when others in the agency have disagreements - Is a stabilizing influence in the agency when dissention occurs [86]
	Cheerleading (viewed as helping [83])	The words and gestures of encouragement and reinforcement of coworkers' accomplishments and professional development - Encourages others when they are down [86]

Sportsmanship (viewed as helping [83])	A willingness to tolerate the inevitable inconvenience and impositions of work without complaining [83] - Consumes a lot of time complaining about trivial matters (R—reverse coded) - Always focuses on what's wrong, rather than the positive side (R) - Tends to make “mountains out of molehills” (R) - Always finds fault with what the organization is doing. (R) - Is the classic “squeaky wheel” that always needs greasing (R) [84]

3. Study design

In this section, we discuss the design of the overall project, including algorithm and software design as well as the design of the proposed domain study, addressing the basic research strategy, concepts to be examined, sample populations and proposed data collection and analysis techniques. We first discuss the goals and general design of the study. We then present the details of each aspect of the study.

Planned system implementation

In the first phase of the project, the focus of the information science research will be on developing algorithms for information extraction, viewed as a token classification problem, and for active learning of classification rules. This development involves determining features of tokens as a basis for learning and developing information extraction and active learning algorithms. The first step will involve delineating the predictable linguistic features on which algorithms to detect the research-relevant concepts can be based. In order to apply token classification to the input text, more information is need than just the individual words. We will apply Syracuse University’s TextTagger system as a preprocessing step to provide additional features for the coding. TextTagger can identify sentence boundaries, part-of-speech tag, stem and lemmatize words, identify various types of phrases, categorize named entities and most common nouns and identify coreferences. These capabilities are generic and the technology is mature: TextTagger has been used in more than 35 projects internally and with other users. Annotated instances will be represented as vectors in the feature space.

For the learning phase, we will utilize MLToolkit, a machine learning and experimentation framework developed by the PI, Dr. Yilmazel, in his doctoral thesis [105]. MLToolkit allows the use of different vector space representations for various categorization problems, and implements selection mechanisms for individual categories. MLToolkit currently includes Support Vector Machines, Decision Trees and Naïve Bayes learning algorithms, as well as several statistical feature selection algorithms. MLToolkit will be extended for this project to add Maximum Entropy and Hidden Markov Model learning algorithms. The experiment management framework in MLToolkit implements various supervised learning experimental designs, such as multi-label categorization, n-fold cross validations and hierarchical categorization. In order to implement the active learning strategies, we will extend the application programming interface (API) of MLToolkit to include the document selection strategies discussed above. Users of the system will be able to control the different learning algorithms, feature space, feature selection and document selection strategies from the web interface.

In phase II, we will integrate the information extraction and active learning algorithms with a user interface in a working prototype system. We plan to implement it as a Web-based system for ease of use (the rule-based system developed in the prior work and described above currently allows examination of the code data via the Web). We will use AJAX (Asynchronous JavaScript and XML) programming to make the user interface for annotation as smooth as possible for the user. Funding is requested for professional programmer support to enable us to create a functioning tool that can be used both for our own work (e.g., on the included domain study) and by other researchers.

Inputs. The system will provide facilities for basic preprocessing of input data, such as conversion from common file formats to text. As well, as a demonstration of the way such a system could be incorporated in scientific cyber-infrastructure, we will provide the capability to retrieve interaction data directly from existing FLOSS data repositories, such as the FLOSSMole project, developed by Crowston as part of prior funded research and currently being extended with support from NSF CNS Grant 07-08437 (with M. Conklin). Features will be extracted from the input documents and use will be able to view and edit the feature space before the system goes into the learning phase.

Processing. The user will select a few of the uploaded documents for initial coding. The system will display the document in its original format as well as the TextTagger annotated version of the same document. Codes can be applied interactively using the system or by importing annotated data from a CAQDAS tool such as Atlas-ti (via its XML export feature). Importing Atlas-ti coded data would also allow calculation of inter-rater reliability, comparing codes from two human coders or between a human coder and the machine coding. From the initial codes the system will infer an initial rule set and use it to process additional documents. The system can ask the user for feedback on the accuracy of the coding during the process and use the newly coded data to refine the rules.

Outputs. Once a satisfactory rule set is obtained, the remaining data will be coded in bulk. The resulting coded data can then be edited for accuracy via the Web interface or in Atlas-ti by a human coder. Data can finally exported for analysis, e.g., as an input to a workflow, or stored back into a data repository for further use. The integration of the coding system with other pieces of scientific cyber-infrastructure will facilitate the use of mixed data analysis, incorporating both quantitative and qualitative data.

Domain study

In this section we will present the design of the domain study in more detail, covering in turn sample selection, data collection and cleaning and data analysis, distinguishing between manual analysis in Phase I and computer-supported analysis in Phase II.

Sample selection. We will start each phase by identifying promising distributed groups for study. During the first phase, we will focus on a small number of groups (on the order of six). In the second phase, the size of the sample will be limited by the available data and processing power (computer and human). In choosing these groups we will apply the previously developed effectiveness assessments (described above) as a theoretical sampling filter to ensure that we have groups of different types with varying degrees of effectiveness. We will also take into consideration some pragmatic considerations, such as selecting only projects where we have access to the needed data. Because of our prior experience in the area, we plan to focus our analysis, at least initially, on FLOSS software development groups (we also have access to interaction data from other kinds of cyber-infrastructure supported collaborations, which can be analyzed time permitting). There are thousands of FLOSS projects, spanning a wide range of applications. Due to their size, success and influence, the Linux operating system and the Apache Web Server and related projects are the most well known, but hundreds of others are in widespread use, including projects on Internet infrastructure (e.g., sendmail, bind), user applications (e.g., Mozilla, OpenOffice) and programming languages (e.g., Perl, Python, gcc) and even enterprise systems (e.g., eGroupware, Compiere, openCRX). Key to our interest is the fact that most FLOSS software is developed by self-organizing distributed groups comprising professionals, users [96-98] and other volunteers working in loosely-coupled groups. These groups are close to pure virtual groups in that developers contribute from around the world, meet face to face infrequently if at all, and coordinate their activity primarily using a cyber-infrastructure [87, 100]. The groups have a high isolation index [77] in that most group members work on their own and in most cases for different organizations (or no organization at all). While these features place FLOSS groups at one end of the continuum of distributed work arrangements, the emphasis on distributed work makes them useful as a research setting for isolating the implications of this organizational innovation. For Phase I, we will chose projects that produce comparable systems in order to control for the nature of the program, thus allowing a more direct comparison of the groups' effectiveness. For example, in the HSD grant described above, we compared Internet Messaging client projects [51].

Data collection and cleaning. To explore the concepts identified in the conceptual development section of this proposal (Table 3), we will collect and analyze a range of data (e.g., e-mail archives, computer logs, primary and secondary project source documents and possibly supplemented with interviews with members of the initial projects). The most voluminous source of data will be collected from archives of CMC tools used to support the groups' interactions [54, 68]. These data are useful because they are unobtrusive measures of the group's behaviors [101]. In particular, mailing list archives will be a primary source of interaction data that illuminates the role of social maintenance, as email is one of the primary tools used to support group communication [67]. In the FLOSS setting, such archives are the primary mode of communication and so contain a huge amount of data (e.g., the Linux kernel list receives 5-7000 messages per month, the Apache httpd list receives an average of 40 messages a day). While in some cases the raw data are already available, significant effort is needed to extract scientifically useful information from them. The initial processing to prepare the data for analysis will be to download the data from the message archives, clean the data (e.g., by removing unnecessary coding from attachments), provide descriptive metadata on each archive, and extract the date, sender and any individual recipients' names, the sender of the original message, in the case of a response, and text of each message. In this preparatory stage, we will record available demographic data such as gender, region, organization and role within the group. We will leverage the work being done as part of one PI's NSF-supported cyberinfrastructure development grant, which is already engaged in capturing and processing email data.

Data analysis. While voluminous, the raw data described above are at a low level of abstraction. The processed data will be analyzed using a variety of techniques to raise the level of conceptualization to fit our theoretical perspective and thus answer our research questions. In phase I, we will use CAQDAS tools for content analysis to develop an initial training set for the new tool. Data will be content analyzed following the process suggested by Miles and Huberman [72], iterating between data collection, data reduction (coding), data display, and drawing and verifying conclusions. A proportion of messages (ideally 100%) will be coded by two individuals to enable calculations of reliability. The initial (deductive) framework will be based on the conceptual development reviewed above, but we plan to evolve this framework based on our experiences with the data. As such, the research will also engage open code instances of group maintenance and other social support mechanisms. We plan to examine the relationship between different aspects of group maintenance and group synergy, as well as other aspects of the group process. In addition, this phase will allow us to develop hypotheses about the relationship between group maintenance behaviour and group performance across various settings, based on a developing understanding of the processes of group maintenance and its role in the life of the groups, hypotheses that can be tested with broader data in Phase II.

System evaluation

A key aspect of the project will be evaluation of the performance of the NLP and ML algorithms and of the overall system for the task of content analysis. We anticipate carrying out evaluations at multiple stages in the projects, gradually increasing their scope. The performance of the NLP and ML algorithms will be evaluated initially by comparing the coding they do to human coding in order to determine precision, recall and utility, as in the previous example. The training data for this purpose will be the data coded as part of this project. Because the group maintenance codebook includes a wide range of types of code, these tests will provide a good test of the generality and limits of the proposed approach. As well, drawing on the results of our prior work, we can test the system using already coded decision-making process data, which will allow a three-way comparison of the performance of automatically learned rules to the hand developed rules to the human coder. Tests with the decision-process data will be limited because we will not want to spend much time doing extensive additional coding, which would be needed to evolve the rules, but having this body of data will allow us to get started immediately, even while the human coding for group maintenance is on-going. Finally, the key evaluation will be how well the system as a whole does at supporting human coders and thus speeding the process of qualitative data analysis. This evaluation will be carried out at the end of the project examining the coding done using the tool and assessing measures such as the speed and volume of coding, the precision of the coding and thus the amount of rework needed and the general capability to support the domain of research.

Management plan

Based on preliminary assessment of the effort required, we are requesting funding for an interdisciplinary team comprising two PIs, one in information science and one in the research domain, two graduate students, and a professional programmer. Both PIs, Drs. Kevin Crowston and Ozgur Yilmazel will work during the summer on project management and research design (1 month), and devote 10% of effort during the academic year to project management and oversight (1/2 day per week). Both PIs will share in project selection, overall project design and report writing. Each PI will be responsible for designing specific aspects of the project and overseeing those aspects:

- Dr. Crowston will direct the project and be responsible for project oversight and reporting and will lead the domain research on the FLOSS groups.
- Dr. Yilmazel will lead the computer/information science research team in NLP tool development and integration.

The graduate students will support the principal investigators in sample selection, definition of constructs and variables, and will have primary responsibility for data collection and analysis, under the oversight of the PIs. As noted above, manual content analysis is extremely labor intensive. However, since we already have an initial codebook for group maintenance, we believe that one student working full time during the first year will be sufficient to support Phase I, with some support from the second student, e.g., to establish inter-coder reliability. To work on development of the NLP and ML algorithms and their integration into a functional data-coding tool we are requesting funding for a research professor and professional programmer, assisted by one student.

For Phase II, the focus of the project will be using the tool to code larger numbers of projects. We have requested some funding for programmers to make any necessary fixes or improvements to the system, but we anticipate that the bulk of the effort will be spent on computer-assisted coding and analysis of the coding process and the coded results. Again, the students will be the primary data coders and we have requested funding for two students full time during the second year. A time line is included as part of the budget justification to show how the requested resources will be employed.

We will employ two main project management techniques. First, we will have regular meetings of the project members to share findings and to plan the work. Initially, these will be every other week, but the frequency of meetings will be adjusted depending on our experience and the pace of the work being carried out at the time. These formal meetings of all project participants will augment the regular interaction of the teams of PIs and students working on the data analysis and expected frequent interactions of the students as they analyze data from the same projects. The NLP development team will meet semi-weekly during the design phases and then weekly during implementation. The experience of this team on the existing toolset bodes well for an accelerated process of iterative requirements, implementation, usage, and new requirements. Second, an initial project activity will be the development of a more detailed timeline (based on the initial one in the budget justification section) against which progress will be measured. The budget includes support during summer and academic year to support these activities.

4. Conclusion

In this proposal, we discussed the challenges of qualitative data analysis and the possibility of using innovative NLP and ML techniques to support it. These techniques will be deployed in a prototype data analysis tool with a human-in-the-loop and active learning. As an example application to prove the utility of the proposed tool, we develop a conceptual framework and a research plan to investigate group maintenance functions within distributed groups. The proposed project will have both intellectual and broader impacts.

Expected intellectual merits

The intellectual merit of the proposed research is three-fold. First, the innovative information science contribution of the proposal is the integration of information extraction and active learning in an interactive system to make practical the use of a system for coding qualitative data in various domains. A validation study will apply the tool to a diverse set of codes, providing evidence of the generality and limits of the approach. Second, the project addresses a fundamental methodological problem in the broad domain of qualitative research, namely dealing with large quantities of unstructured qualitative data, by applying innovative information extraction technologies. Finally, a second domain science contribution of the study is to address a fundamental problem in the application domain of organizational behaviour, namely group maintenance in a novel setting, namely distributed groups working together using cyber-infrastructure, to advance our understanding of the effects of interpersonal relationships on the functioning, effectiveness and innovation of groups who rely on innovative applications of computer-mediated communications (CMC).

Expected broader impacts

The project has numerous broader impacts. In addition to the expected intellectual contributions described above, the proposed research will benefit society by providing a component of useful infrastructure for qualitative science research, thus contributing to the infrastructure of science. In particular, we will integrate the tool with cyber-infrastructure currently being developed by one of the PIs with other NSF support. A second goal is to provide generalizable knowledge to improve the effectiveness of distributed groups, a further benefit to society. Such groups are an increasingly important approach to work such as software development, scientific research and policy development. Distributed work is potentially transformative for organization but the separation between members of distributed groups creates difficulties in building social relations, which may ultimately result in a failure of the group to be effective. For the potential of distributed groups to be realized, research is needed on how to make them engaging and motivating to members. Understanding the role of group maintenance in these settings and the relation to group performance will help us develop guidelines to improve performance and foster innovation.

To ensure that our study has a significant impact, we plan to broadly disseminate results through journal publications, conferences, workshops and on our Web pages. We plan to work to integrate the prototype tool into the cyberinfrastructure being developed and to make it available to qualitative researchers (though supporting such distribution has some challenges that will have to be faced in the future). The findings about distributed teams can be disseminated as well as through our interaction with the leaders and members of distributed teams. These results could also be incorporated into the curricula of the professional degrees of the Syracuse University School of Information Studies, as well as improving the pedagogy of our courses and degree programs, as these programs are offered on-line and thus involve distributed groups. Finally, the project will promote teaching, training, and learning by students in the research project, providing them the opportunity to develop skills in model development, theory application, data collection and analysis.

Results from prior NSF funding

One of the PIs for this grant, Crowston, has been funded by several additional NSF grants within the past 48 months, including the grant reviewed above, HSD 05–27457 ($684,882, 2005–2008, with R. Heckman, E. Liddy and N. McCracken), *Investigating the Dynamics of Free/Libre Open Source Software Development Teams*. Crowston also received funding for IIS 04–14468 ($327,026, 2004–2006) and SGER IIS 03–41475 ($12,052, 2003–2004), both entitled *Effective work practices for Open Source Software development* and for NSF CNS Grant 07-08437 ($200,000, 2007–2010, with M. Conklin, Elon), for *Collaborative Research: CRI: CRD: Data and analysis archive for research on Free and Open Source Software and its development*. The first three of these grants have supported a study of the evolution of effective work practices for distributed groups, specifically, for teams of free/libre open source software (FLOSS) developers. The funding enabled travel to conferences (e.g., *ApacheCon* and *OSCon*) to observe and interview FLOSS developers and to present preliminary results, and for the purchase of data analysis software and equipment. The final grant is supporting the development of cyber-infrastructure to support the FLOSS research community more broadly. We plan to leverage this investment in supporting the proposed work. Overall, this work has resulted in nine journal papers [21-23, 25, 27, 28, 30, 32, 58] with others under review [1], multiple conference papers [e.g., 15, 19, 24, 29, 31, 33, 50, 60, 61, 69, 93] and workshop presentations [14, 16, 17, 20, 59]. These grants have supported a total of six PhD students; several others have been involved in specific aspects of the work. The HSD grant included a component applying NLP techniques to analyze large corpora of email (as noted above) and provided significant experience working in an interdisciplinary team.

Crowston's final grant is IIS 04–14482 ($302,685, 2005–2006, with Barbara Kwasnik), for *How can document-genre metadata improve information-access for large digital collections?* The grant partially supported work on several publications [e.g., 64] conference papers, a conference mini-track and journal special issue [66]. Earlier work by the PIs on genre has appeared in journals [e.g., 26] and conference papers [e.g., 65]. As well, we have developed a classification of document genre for Web documents and a corpus of more than 2500 Web pages coded by genre, which will be used to support further experiments. The grant has funded two PhD students; two others are involved in aspects of the research.

References

[1] E. E. Allen, R. Heckman, Q. Li, U. Y. Eseryel, K. Crowston, J. Howison, and K. Wei, “Natural language processing (NLP) tools for the content coding of data,” *Journal of Information Technology and Politics*, Special issue on Text Annotation for Political Science Research, Under review.

[2] T. Anderson, L. Rourke, W. Archer, and R. Garrison, “Assessing teaching presence in computer conferencing transcripts,” *Journal of the Asynchronous Learning Network*, vol. 5, no. 2, 2001.

[3] R. Aviv, “Educational performance of ALN via content analysis,” *The Journal Of Asynchronous Learning Networks*, vol. 4, no. 2, 2000.

[4] R. Axelrod, *The evolution of co-operation*. New York: Basic Books, 1985.

[5] R. F. Bales, “A set of categories for the analysis of small group interaction,” *American Sociological Review*, vol. 15, no. 2, pp. 257–263, 1950.

[6] M. Beißwenger, “Bibliography of Chat Communications,” [Online document], 2003, [cited 17 February 2004], Available http://www.chat-bibliography.de/

[7] E. Bender, “Rules of the Collaboratory Game,” in *Technology Review*, vol. 23 November, 2004.

[8] S. Booth, “Computer-assisted analysis in qualitative research,” *Computers in Human Behaviour*, vol. 9, pp. 203-211, 1993.

[9] B. R. Boyce, C. T. Meadow, and D. H. Kraft, *Measurement in information science*. San Diego: Academic Press, 1994.

[10] P. Brown and S. Levinson, *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press, 1987.

[11] J. W. Carey, P. H. Wenzel, C. Reilly, J. Sheridan, and J. M. Steinberg, “CDC EZ-Text: Software for management and analysis of semi-structured qualitative data sets,” *Cultural Anthropology Methods*, vol. 10, pp. 14–20, 1998.

[12] H. L. Chieu and H. T. Ng, “A maximum entropy approach to information extraction from semi-structured and free text,” in *Eighteenth national conference on Artificial intelligence*. Edmonton, Alberta, Canada, 2002, pp. 786–791

[13] D. L. Cogburn, “Diversity matters, even at a distance: Evaluating the impact of computer-mediated communication on civil society participation in the World Summit on the Information Society,” *Information Technologies and International Development*, vol. 1, no. 3, pp. 15–42, 2004.

[14] M. Conklin, J. Howison, and K. Crowston, “Collaboration Using OSSmole: A repository of FLOSS data and analyses,” in *Symposium on Mining Software Repositories*. St. Louis, 2005.

[15] K. Crowston, H. Annabi, and J. Howison, “Defining Open Source Software project success,” presented at Proceedings of the 24th International Conference on Information Systems (ICIS 2003), Seattle, WA, 2003.

[16] K. Crowston, H. Annabi, J. Howison, and C. Masango, “Effective work practices for Software Engineering: Free/Libre Open Source Software Development,” presented at WISER Workshop on Interdisciplinary Software Engineering Research, SIGSOFT 2004/FSE-12 Conference,, Newport Beach, CA, 2004.

[17] K. Crowston, H. Annabi, J. Howison, and C. Masango, “Towards a portfolio of FLOSS project success measures,” presented at Workshop on Open Source Software Engineering, 26th International Conference on Software Engineering, Edinburgh, 2004.

[18] K. Crowston, H. Annabi, J. Howison, and C. Masango, “Effective work practices for FLOSS development: A model and propositions,” in *Proceedings of the Hawai’i International Conference on System Science (HICSS)*. Big Island, Hawai'i, 2005.

[19] K. Crowston, R. Heckman, H. Annabi, and C. Masango, “A structurational perspective on leadership in Free/Libre Open Source Software teams,” presented at OSSCon, Genova, Italy, 2005.

[20] K. Crowston and J. Howison, “The social structure of Open Source Software development teams,” presented at The IFIP 8.2 Working Group on Information Systems in Organizations Organizations and Society in Information Systems (OASIS) 2003 Workshop, Seattle, WA, 2003.

[21] K. Crowston and J. Howison, “The social structure of free and open source software development,” *First Monday*, vol. 10, no. 2, 2005.

[22] K. Crowston and J. Howison, “Hierarchy and centralization in Free and Open Source Software team communications,” *Knowledge, Technology & Policy*, vol. 18, no. 4, pp. 65–85, 2006.

[23] K. Crowston, J. Howison, and H. Annabi, “Information systems success in Free and Open Source Software development: Theory and measures,” *Software Process—Improvement and Practice*, vol. 11, no. 2, pp. 123–148, 2006.

[24] K. Crowston, J. Howison, C. Masango, and U. Y. Eseryel, “Face-to-face interactions in self-organizing distributed teams,” presented at Academy of Management Conference, Honolulu, HI, 2005.

[25] K. Crowston, J. Howison, C. Masango, and U. Y. Eseryel, “The role of face-to-face meetings in technology-supported self-organizing distributed teams ” *IEEE Transactions on Professional Communications*, no. September, 2007.

[26] K. Crowston and B. H. Kwasnik, “Can document-genre metadata improve information access to large digital collections?,” *Library Trends*, vol. 52, no. 2, pp. 345–361, 2003.

[27] K. Crowston, Q. Li, K. Wei, U. Y. Eseryel, and J. Howison, “Self-organization of teams for free/libre open source software development,” *Information and Software Technology*, vol. 49, no. 6, pp. 564–575, 2007.

[28] K. Crowston and B. Scozzi, “Open source software projects as virtual organizations: Competency rallying for software development,” *IEE Proceedings Software*, vol. 149, no. 1, pp. 3–17, 2002.

[29] K. Crowston and B. Scozzi, “Coordination practices for bug fixing within FLOSS development teams ” presented at Presentation at 1st International Workshop on Computer Supported Activity Coordination, 6th International Conference on Enterprise Information Systems, Porto, Portugal, 2004.

[30] K. Crowston and B. Scozzi, “Coordination practices within Free/Libre Open Source Software development teams: The bug fixing process,” *Journal of Database Management*, vol. Special Issue on Open Source Software, In press.

[31] K. Crowston, K. Wei, Q. Li, U. Y. Eseryel, and J. Howison, “Coordination of Free/Libre Open Source Software development,” presented at International Conference on Information Systems (ICIS 2005), Las Vegas, NV, USA, 2005.

[32] K. Crowston, K. Wei, Q. Li, U. Y. Eseryel, and J. Howison, “Self-organization of teams in free/libre open source software development,” *Information and Software Technology Journal, Special issue on Understanding the Social Side of Software Engineering, Qualitative Software Engineering Research*, vol. 49, pp. 564–575, 2007.

[33] K. Crowston, K. Wei, Q. Li, and J. Howison, “Core and periphery in Free/Libre and Open Source software team communications,” presented at Hawai'i International Conference on System System (HICSS-39), Kaua'i, Hawai'i, 2006.

[34] M. Daniel, “John Dewey and Matthew Lipman: Pragmatism in education,” in *Children: Thinking and Philosophy*, D. Camhy, Ed. Sankt Augustin: Academia Verlag, 1994.

[35] A. Finn and N. Kushmerick, “Active learning selection strategies for information extraction,” in *Proceedings of the International Workshop on Adaptive Text Extraction and Mining*, 2003.

[36] A. Finn and N. Kushmerick, “Learning to classify documents according to genre,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1506–1518, 2006.

[37] P. Galison and B. Hevly, *Big science: The growth of large-scale research*. Palo Alto, CA: Stanford University, 1992.

[38] R. Garrison, T. Anderson, and W. Archer, "Critical thinking in a text-based environment: Computer conferencing in higher education," *The Internet and Higher Education*, vol. 2, no. 2-3, pp. 87-105, 2000.

[39] R. Garrison, T. Anderson, and W. Archer, "Critical thinking, cognitive presence and computer conferencing in distance education," *American Journal of Distance Education*, vol. 15, no. 1, pp. 7-23, 2001.

[40] D. L. Gladstein, "Groups in context: A model of task group effectiveness," *Administrative Science Quarterly*, vol. 29, no. 4, pp. 499–517, 1984.

[41] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Publishing, 1967.

[42] E. Goffman, *Interaction ritual: Essays on face-to-face behaviour*. Garden City, NY: Anchor Books, 1967.

[43] D. S. Gouran, C. Brown, and D. R. Henry, "Behavioral correlates of perceptions of quality in decision-making discussions," *Communication Monographs*, no. 45, pp. 51-63, 1978.

[44] J. W. Graham, "Principled organizational dissent: A theoretical essay," in *Research in Organizational Behaviour* vol. 8, B. M. Staw and L. L. Cummings, Eds. Greenwich, CT: JAI Press, 1986, pp. 1-52.

[45] P. Gronn, "Distributed leadership as a unit of analysis," *The Leadership Quarterly*, vol. 13, no. 4, pp. 423–451, 2002.

[46] J. R. Hackman, "The design of work teams," in *The Handbook of Organizational Behaviour*, J. W. Lorsch, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1987, pp. 315–342.

[47] M. Handel and J. D. Herbsleb, "What is chat doing in the workplace? ," in *Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW)*. New Orleans, LA, 2002.

[48] C. Hass, *Writing technology: Studies on the materiality of literacy*. Manwah, N.J.: Erlbaum, 1996.

[49] R. Heckman and H. Annabi, "A content analytic comparison of FTF and ALN case-study discussions," in *36th Annual Hawaii International Conference on System Sciences (HICSS'03)*. Big Island, Hawaii: IEEE Press, 2003.

[50] R. Heckman, K. Crowston, U. Y. Eseryel, J. Howison, E. Allen, and Q. Li, "Emergent decision-making practices In Free/Libre Open Source Software (FLOSS) development teams," in *3rd International Conference on Open Source Software*. Limerick, Ireland, 2007.

[51] R. Heckman, K. Crowston, Q. Li, E. Allen, Y. Eseryel, J. Howison, and K. Wei, "Emergent decision-making practices in technology-supported self-organizing distributed teams," in *Proceedings of the International Conference on Information Systems (ICIS 2006)*. Milwaukee, WI, 10–13 Dec, 2006.

[52] D. Hellriegel, S. E. Jackson, J. Slocum, and G. Staude, *Management*. Cape Town: OUP, 2001.

[53] J. D. Herbsleb and A. Mockus, "An empirical study of speed and communication in globally-distributed software development," *IEEE Transactions on Software Engineering*, vol. 29, no. 3, pp. 1–14, 2003.

[54] J. D. Herbsleb, A. Mockus, T. A. Finholt, and R. E. Grinter, "An empirical study of global software development: Distance and speed," presented at Proceedings of the International Conference on Software Engineering (ICSE 2001), Toronto, Canada, 2001.

[55] S. C. Herring, "Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives." Philadelphia: John Benjamins, 1996.

[56] S. R. Hiltz and M. Turoff, *The Network Nation*. Cambrige, MA: MIT Press, 1993.

[57] R. J. House and R. N. Aditya, "The social scientific study of leadership: Quo vadis?," *Journal of Management*, vol. 23, no. 3, pp. 409–473, 1997.

[58] J. Howison, M. Conklin, and K. Crowston, "FLOSSmole: A collaborative repository for FLOSS research data and analyses," *International Journal of Information Technology and Web Engineering*, vol. 1, no. 3, pp. 17–26, 2006.

[59] J. Howison, M. S. Conklin, and K. Crowston, “OSSmole: A collaborative repository for FLOSS research data and analyses,” presented at 1st International Conference on Open Source Software, Genova, Italy, 2005.

[60] J. Howison and K. Crowston, “The perils and pitfalls of mining SourceForge,” presented at Presentation at the Workshop on Mining Software Repositories, 26th International Conference on Software Engineering, Edinburgh, Scotland, 2004.

[61] J. Howison, K. Inoue, and K. Crowston, “Social dynamics of FLOSS team communications,” presented at The Second International Conference on Open Source Systems, Como, Italy, 2006.

[62] S. L. Jarvenpaa and D. E. Leidner, “Communication and trust in global virtual teams,” *Organization Science*, vol. 10, no. 6, pp. 791–815, 1999.

[63] M. Konovsky and D. Organ, “Dispositional and contextual determinants of organizational citizenship behaviour,” *Journal of Organizational Behaviour*, vol. 17, no. 3, pp. 253-266, 1996.

[64] B. H. Kwasnik, Y.-L. Chun, K. Crowston, J. D’Ignazio, and J. Rubleske, “Challenges in creating a taxonomy for genres of digital documents,” in *Proceedings of the 2006 ISKO Conference*. Vienna, Austria, 2006.

[65] B. H. Kwasnik and K. Crowston, “A framework for creating a facetted classification for genres: Addressing issues of multidimensionality,” in *Proceedings of the Hawai'i International Conference on System Science (HICSS)*. Big Island, Hawai'i, 2004.

[66] B. H. Kwasnik and K. Crowston, “Genres of digital documents: Introduction to the special issue,” *Information, Technology & People*, vol. 18, no. 2, pp. 76–88, 2005.

[67] G. F. Lanzara and M. Morner, “Making and sharing knowledge at electronic crossroads: the evolutionary ecology of open source,” presented at 5th European Conference on Organizational Knowledge, Learning and Capabilities, Innsbruck, Austria, 2004.

[68] G. K. Lee and R. E. Cole, “From a firm-based to a community-based model of knowledge creation: The case of Linux kernel development,” *Organization Science*, vol. 14, no. 6, pp. 633–649, 2003.

[69] Q. Li, K. Crowston, R. Heckman, and J. Howison, “Language and power in self-organizing distributed teams,” presented at OCIS Division, Academy of Management Conference, Atlanta, GA, 2006.

[70] M. Lipman, “Philosophy for Children,” *Metaphilosophy*, vol. 7, pp. 17-39, 1976.

[71] L. L. Martins, L. L. Gilson, and M. T. Maynard, “Virtual teams: What do we know and where do we go from here?,” *Journal of Management*, vol. 30, no. 6, pp. 805-835, 2004.

[72] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook*, 2nd ed. Thousand Oaks: Sage Publications, 1994.

[73] D. A. Morand and R. J. Ocker, “Politeness theory and Computer-Mediated communication: A Sociolinguistic Approach to Analyzing Relational Messages,” in *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS’03)*, 2003.

[74] M. Morison and J. Moir, “The role of computer software in the analysis of qualitative data: Efficient clerk, research assistant or Trojan horse?,” *Journal of Advanced Nursing*, vol. 28, no. 1, pp. 106–116, 1998.

[75] M. D. Myers, “Qualitative research in information systems,” *MIS Quarterly*, vol. 21, no. 2, pp. MISQ Discovery updated version 28 December 1999, http://www.auckland.ac.nz/msis/isworld/, page accessed on 31 December 1999, 1997.

[76] D. R. Newman, C. Johnson, C. Cochrane, and B. Webb, “An experiment in group learning technology: evaluating critical thinking in face-to-face and computer-supported seminars,” *Interpersonal Computer and Technology Journal*, vol. 4, no. 1, pp. 57-74, 1996.

[77] M. O’Leary and J. Cummings, “The Spatial, Temporal, and Configurational Characteristics of Geographic Dispersion in Teams,” presented at Academy of Management Conference, Denver, CO, 2002.

[78] R. J. Ocker, “The Relationship between Interaction, Group Development, and Outcome: A Study of Virtual Communication,” in *Proceedings of the Thirty-forth Annual Hawaii International Conference on System Science*, 2001.

[79] R. Ohlsson, “An early form of the community of inquiry: The study circle,” *Thinking*, vol. 14, no. 27–28, 1998.

[80] D. R. Olson, *The world on paper: The conceptual and cognitive implications of reading and writing* Cambridge and New York: Cambridge University Press, 1994.

[81] D. Organ, *Organizational citizenship behaviour: The good soldier syndrome*. Lexington, MA: Lexington Books, 1988.

[82] D. Organ, P. Podsakoff, and S. MacKenzie, *Organizational citizenship behaviour: Its nature, antecedents, and consequences*. Thousand Oaks, CA: SAGE Publications, 2006.

[83] D. W. Organ, P. M. Podsakoff, and S. B. MacKenzie, *Organizational citizenship behaviour: Its nature, antecedents, and consequences*. Thousand Oaks, CA: SAGE Publications, 2006.

[84] P. M. Padsakoff, S. B. MacKenzie, R. H. Mooreman, and R. Fetter, “Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors,” *Leadership Quarterly*, vol. 1, pp. 107-142, 1990.

[85] C. L. Pearce and J. A. Conger, “All those years ago: The historical underpinnings of shared leadership,” in *Shared Leadership: Reframing the Hows and Whys of Leadership* C. L. Pearce and J. A. Conger, Eds. Thousand Oaks, CA: Sage, 2003, pp. 1–18.

[86] P. M. Podsakoff and S. B. MacKenzie, “Organizational citizenship behaviors and sales unit effectiveness,” *Journal of Marketing Research*, vol. 31, pp. 351-363, 1994.

[87] E. S. Raymond, “The cathedral and the bazaar,” *First Monday*, vol. 3, no. 3, 1998.

[88] M. Ridley, *The Origins of Virtue: Human Instincts and the Evolution of Cooperation*. New York: Viking, 1996.

[89] T. B. Roby, “The executive function in small groups,” in *Leadership and Interpersonal Behaviour*, L. Petrullo and B. M. Bass, Eds. New York: Holt, Rinehart and Wilson, 1961.

[90] L. Rourke, T. Anderson, D. R. Garrison, and W. Archer, “Assessing social presence in asynchronous, text-based computer conferencing,” *Journal of Distance Education*, vol. 14, no. 2, pp. 51–70, 1999.

[91] J. Scheffer, “Data mining in the survey setting: Why do children go off the rails?,” *Res. Lett. Inform. Math. Sci.*, vol. 3, pp. 161–189, 2002.

[92] W. C. Schutz, “The ego, RIFO theory, and the leader as completer,” in *Leadership and Interpersonal Behaviour*, L. Petrullo and B. M. Bass, Eds. New York: Holt, Rinehart and Wilson, 1961, pp. 48–65.

[93] B. Scozzi, K. Crowston, U. Y. Eseryel, and Q. Li, “Shared mental models among open source software developers,” in *Hawai'i International Conference on System Science*. Big Island, Hawai'i, 2008.

[94] C. Smith, D. Organ, and J. Near, “Organizational citizenship behaviour: Its nature and antecedents,” *Journal of Applied Psychology*, no. 68, pp. 653-663, 1983.

[95] V. Vapnik, *The Nature of Statistical Learning Theory*: Springer-Verlag, 1995.

[96] E. von Hippel, “Innovation by user communities: Learning from open-source software,” *Sloan Management Review*, no. Summer, pp. 82–86, 2001.

[97] E. von Hippel and G. von Krogh, “Exploring the Open Source Software Phenomenon: Issues for Organization Science,” Sloan School of Management, MIT, Cambridge, MA 2002.

[98] E. von Hippel and G. von Krogh, “Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science,” *Organization Science*, vol. 14, no. 2, pp. 209–213, 2003.

[99] R. E. Walton and J. R. Hackman, “Groups under contrasting management strategies,” in *Designing Effective Work Groups*, P. S. Goodman and Associates, Eds. San Francisco, CA: Jossey-Bass, 1986, pp. 168–201.

[100] P. Wayner, *Free For All*. New York: HarperCollins, 2000.

[101] E. Webb and K. E. Weick, "Unobtrusive measures in organizational theory: A reminder," *Administrative Science Quarterly*, vol. 24, no. 4, pp. 650–659, 1979.

[102] E. Welsh, "Dealing with data: Using NVivo in the qualitative data analysis process," *Forum Qualitative Sozialforschung*, vol. 3, no. 2, 2002.

[103] E. M. White, "Assessing higher-order thinking and communication skills in college graduates through writing," *The Journal of General Education*, no. 42, pp. 105-122, 1993

[104] L. Williams and S. Anderson, "Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors," *Journal of Management*, no. 17, pp. 601-617, 1991.

[105] O. Yilmazel, "Empirical Selection of NLP-Driven Document Representations for Text Categorization," PhD Thesis, Syracuse University, Syracuse, NY 2006.

[106] G. Yukl, *Leadership In Organizations*, 5th ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[107] G. Yukl, A. Gordon, and T. Taber, "A hierarchical taxonomy of leadership behaviour: Integrating a half century of behaviour research," *Journal of Leadership & Organization Studies*, vol. 9, no. 1, pp. 15-32, 2002.

Kevin Crowston

Curriculum Vitae

Education

1980–1984 A. B., *magna cum laude,* June 1984, Applied Mathematics (Computer Science), Harvard University.

1984–1991 Ph. D., January 1991, Information Technologies, Sloan School of Management, Massachusetts Institute of Technology.

Appointments

1991–1996 Assistant Professor of Computer and Information Systems,
School of Business, University of Michigan.

1996– Professor of Information Studies (Assistant 1996–2001, Associate 2001–2006),
School of Information Studies, Syracuse University.

Publications (from a total of 53 peer reviewed journal and conference papers)

1. Crowston, K. & Howison, J. (2006). Hierarchy and centralization in Free and Open Source Software team communications. ***Knowledge, Technology & Policy***. *18*(4), 65–85.
2. Crowston, K., Howison, J. & Annabi, H. (2006). Information systems success in free and open source software development: Theory and measures. ***Software Process—Improvement and Practice (SPIP)***, *11*(2), 123–148.
3. Howison, J., Conklin, M. & Crowston, K. (2006). FLOSSmole: A collaborative repository for FLOSS research data and analyses. ***International Journal of Information Technology and Web Engineering (IJITWE)***, *1*(3), 17–26.
4. Crowston, K. and Scozzi, B. (2002). Open source software projects as virtual organizations: Competency rallying for software development. ***IEE Proceedings Software***, 149(1), 3–17.
5. Crowston, K. and Kammerer, E. E. (1998). Coordination and collective mind in software requirements analysis. ***IBM Systems Journal***, 37(2), 227–245.

Other significant publications

1. Crowston, K. & Myers, M. D. (2004). Information technology and the transformation of industries: Three research perspectives. ***Journal of Strategic Information Systems***, *13*(1), 5–28.
2. Watson-Manheim, M.-B., Chudoba, K. M. and Crowston, K. (2002). Discontinuities and continuities: A new way to understand virtual work. ***Information, Technology and People***, *15*(3), 191–209.
3. Crowston, K., Sawyer, S. and Wigand, R. (2001). Investigating the interplay between structure and technology in the real estate industry. ***Information, Technology and People***, *14*(2), 163–183.
4. Malone, T. W., Crowston, K., Lee, J., Pentland, B., Dellarocas, C., Wyner, G., Quimby, J., Osborne, C., Bernstein, A., Herman, G., Klein, M. and O'Donnell, E. (1999). Tools for inventing organizations: Toward a handbook of organizational processes. ***Management Science***, 43(3), 425–443.

5. Crowston, K. (1997). A coordination theory approach to organizational process design. ***Organization Science***, 8(2), 157–175.

Synergistic activities

1. **Maintainer** of ISWorld Website on Information-Related Doctoral Programs, http://isphd.syr.edu/
2. **Invited participant**, Schloss Dagstuhl Perspectives Seminar 04051: "Empirical Theory and the Science of Software Engineering", 25–29 January 2004.
3. **Co-program chair**, IFIP Working Group 8.2 Working Conference on *Virtuality and Virtualization*, Portland, OR, July 2007.

Collaborators in the past 48 months

Eileen Allen (Syracuse)
Hala Annabi (Ohio)
Kathy Chudoba (Florida State)
You-Lee Chun (Syracuse)
Megan Conklin (Elon)
John D'Ignazio (Syracuse)
U. Yeliz Eseryel (Syracuse)
Claudio Garavelli (Polytechnic of Bari)
Robert Heckman (Syracuse)
James Howison (Syracuse)
Carina Ihlström (Halmstad)
Bernhard Katzy (UniBW Munich)
Barbara Kwasnik (Syracuse)
Chei Sian Lee (National University of Singapore)
Qing Li (Syracuse)
Elizabeth D. Liddy (Syracuse)
Chengetai Masango (Syracuse)
Nelson Massad (Florida Atlantic)
Nora Misiolek (Marist College)
Michael Myers (Auckland)
Dmitri Roussinov (Arizona State)
Joseph Rubleske (Syracuse)
Steve Sawyer (Penn State)
Barbara Scozzi (Polytechnic of Bari)
Sandra Sieber (IESE)
Mary-Beth Watson-Manheim (Illinois Chicago)
Kangning Wei (Syracuse)
Rolf Wigand (Arkansas)
Eleanor Wynn (Intel)

Thesis advisors:

Professor Thomas W. Malone (Chair), Deborah Ancona and John Carroll (all of the Sloan School of Management, Massachusetts Institute of Technology).

Thesis advisees (5 current advisees and 2 graduates)

Marcel Allbritton (consultant), Naybell Hernandez, Chengetai Masango, Kangning Wei, James Howison, Qing Li (all of the School of Information Studies, Syracuse University); Hala Annabi (Ohio)

Ozgur Yilmazel

Professional Preparation

Syracuse University	Electrical Engineering Ph.D.	2006
Syracuse University	Electrical Engineering M.Sc.	2002
Osmangazi University	Electrical Engineering B.Sc.	1996

Appointments

2007-	Assistant Research Professor, School of Information Studies, Syracuse University
1999-	Chief Software Engineer, Center for Natural Language Processing, Syracuse University
1998-99	Software Engineer, Syracuse Language Systems, Syracuse, NY
1996-97	Research Assistant, Computer Engineering, Osmangazi University, Eskisehir, Turkey
1993-95	Systems Engineer, Mikrokom Computer Company, Istanbul, Turkey

Related Publications

Yilmazel, O., Symonenko, S., Balasubramanian, N., Liddy, E.D. (2006) Leveraging One-class SVM in Semantic Analysis for Anamalous Content Detection. To appear in Reid, E. (Ed) Terrorism Infomatics

Yilmazel, O. Symonenko, S., Balasubramanian, N., Liddy, E.D. (2005).Improved Document Representation for Classification Tasks for the Intelligence Community. In the Proceedings of the 2005 AAAI Spring Symposium Series.

Yilmazel, O., Balasubramanian, N., Harwell , S.C. , Bailey, J., Diekema, A.R., Liddy, E.D., (2006) Text Categorization for Aligning Educational Standards. To appear in the Proceedings of the HICSS 2007, Hawaii .

Diekema, A.R., Yilmazel, O., Bailey, J., Harwell, S.C.,Liddy, E.D. (2007) Standards Alignment for Metadata Assignment. Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries. Vancouver, British Columbia, Canada, June 18-23, 2007. pp. 398-399.

Yilmazel, O., Finneran, C.M. & Liddy, E.D. (2004). MetaExtract: An NLP System to Automatically Assign Metadata. Proceedings of the 2004 Joint Conference on Digital Libraries .

Other Publications

Ingersoll, G.I, Yilmazel, O. Liddy, E.D. (2006) Finding Questions, 2007 IEEE 23rd International Conference on Data Engineering, ICDE'07, Istanbul.

Yilmazel, O. Symonenko, S., Balasubramanian, N., Liddy, E.D. (2005). Leveraging One-Class SVM and Semantic Analysis to Detect Anomalous Content. ISI 2005 381-388

Liddy, E.D., Diekema, A.R., & Yilmazel, O. (2004). Context-Based Question-Answering Evaluation. In Proceedings of the 27th Annual ACM-SIGIR Conference. Sheffield, England.

Symonenko, S., Liddy, E.D., Yilmazel, O., DelZoppo, R., Brown, E. & Downey, M. (In Press). Semantic Analysis for Monitoring Insider Threats. In Proceedings of 2nd Symposium on Intelligence and Security Informatics. Tucson, Arizona.

Liddy, E.D., Diekema, A.R., Yilmazel, O., Chen, J., Harwell, S., and He, L. (2004) Finding Answers to Complex Questions. In Maybury, M. (Ed.) New Directions in Question Answering. AAAI-MIT Press.

Synergistic Activities

- Developed MLToolkit: a flexible machine learning and experiment management system for categorization problems.
- Active participant in open-source projects:
  - Lucene – Open-source search software
  - Zemberek – Platform independent, general-purpose Natural Language Processing library and toolset designed for Turkish.
- Visiting Faculty in Anadolu University, Turkey – Taught a course on Applied Software Engineering.

Selected Recent Research Grants Received

2006-2007	Disruptive Technologies Office
2006-2009	National Institute of Medicine
2005-2006	Syracuse Research Corporation
2003-2006	NASA
2004-2005	Syracuse Research Corp.
2003-2005	MySentient, Boulder, Co.
2002-2004	National Science Digital Library Project.
2002-2003	Syracuse Research Corp.
2001-2004	DARPA
2000-2001	NASA
1999-2000	In-Q-Tel

Collaborators & Other Affiliations

Gay, Gerri – Cornell University
Sutton, Stuart – University of Washington
Ingraffia, Tony – Cornell University
Turner, Anne – University of Washington
Merrill, Jacqueline – Columbia University
DelZoppo, Robert – Syracuse Research Corp.

Graduate Advisors

Can Isik – Syracuse University
Elizabeth D. Liddy – Syracuse University

Budget Justification

A. Salaries and Wages – Senior Personnel

The PI, Dr. Kevin Crowston, will work during the summer (1 month/year) on sample selection, detailed project design, integration of data analysis and publication of results. Dr. Crowston will be responsible for overall project direction and coordination, for assuring successful project completion, including submission of NSF progress reports, as required. Dr. Yilmazel, a research professor, will work on NLP and ML algorithm design and will direct the programming team developing the system. Funding is requested for 3.6 months of Dr. Yilazel's time in year 1 and 1.8 months in year 2. The PIs will jointly be responsible for the review of the data and preparation of manuscripts for publication.

B. Salaries and Wages – Other Personnel

Funding is requested for two Ph.D. students, 50% academic year and 100% summer effort, for a total of 2200 hours/year (4400 hours in two years). The graduate students will support the principal investigators in sample section and will have primary responsibility for data analysis, under the oversight of the PIs. Additional funding is requested for programmer staff, half time during year 1 and 20% effort during year 2. Programmer time will be critical to achieve the desired level of functionality of the research prototype to be able to use it in production.

C. Fringe Benefits

Fringe Benefits are calculated as direct costs in accordance with Syracuse University's indirect cost rate agreement (Department of Health and Human Services, 17.0% for faculty during the summer, 17.2% for graduate students, 32.4% for staff). Actual rates in place at the time of an award will be charged.

E2. Travel:

Travel support is requested for students and PIs to disseminate results at academic conferences (one trip each, $1500/trip).

G. Other Direct Costs

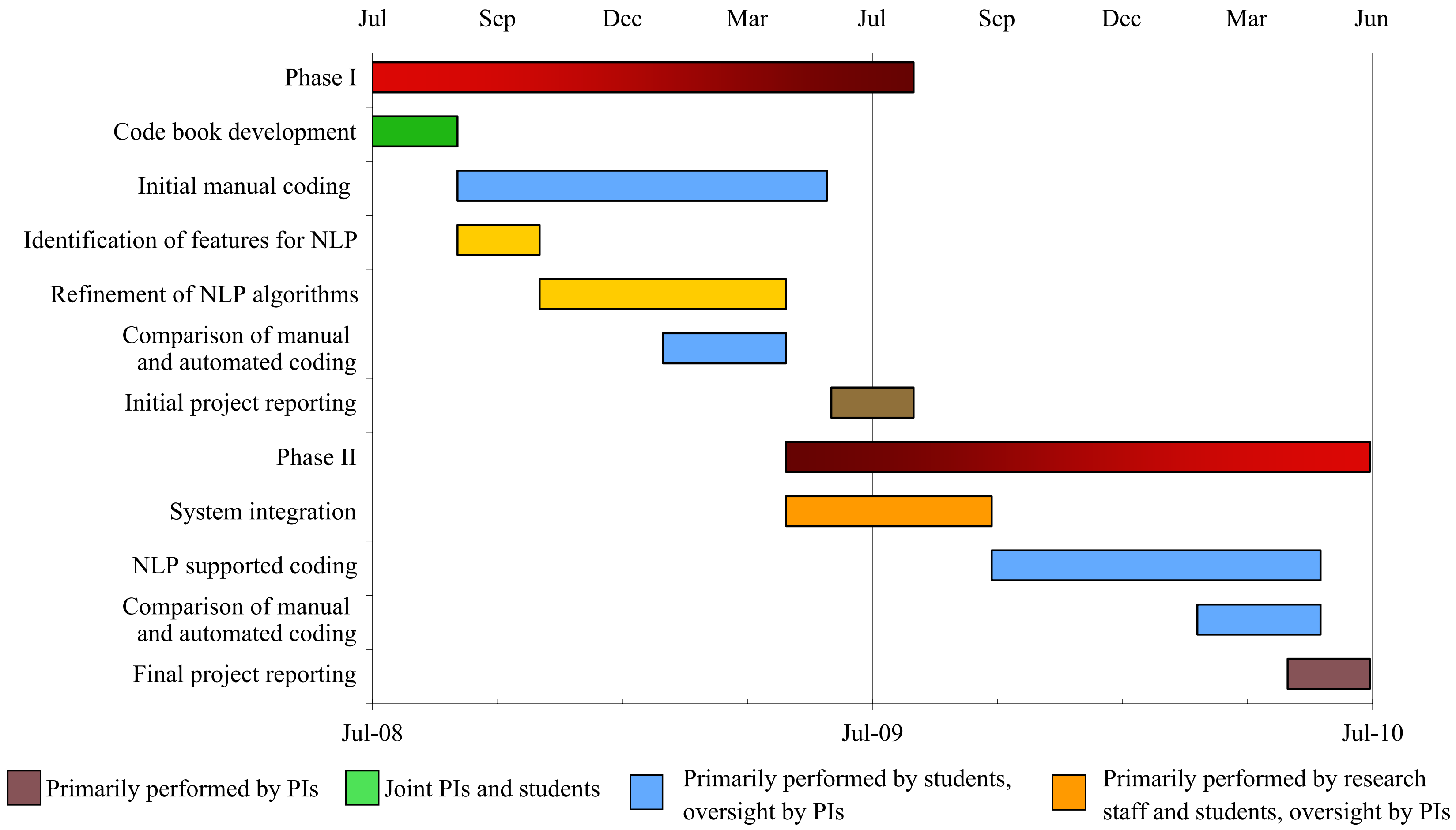
Tuition

A total of $53,352 is requested for partial support of tuition for two graduate students (12 credit hours per year per student at $1,079/credit for Year 1 and $1,144/credit for Year 2).

I. Indirect Costs

Indirect Costs are calculated in accordance with Syracuse University's federally negotiated indirect cost rate agreement (Department of Health and Human Services), which is currently 46% of modified total direct costs (MTDC).

The following proposed timeline for the project indicates how the requested resources will be applied.



Facilities, equipment and other resources

Syracuse University is one of the largest and most comprehensive independent universities in the United States. Founded in 1870, Syracuse offers excellent facilities, equipment and other resources for research and study in many academic and professional disciplines.

The **School of Information Studies** is a leading center for innovative programs in information policy, information behavior, information management, information systems, information technology and information services. Its approach stands out from other institutions that offer computer science, management, information science and related programs in that our focus is on users and user information needs as a starting point for integrating information and information technology into organizations. The faculty of the School crosses disciplinary boundaries to integrate the common elements of information management in business, government, education, and nonprofit settings, including the relationship of information and knowledge, electronic and traditional libraries, information systems and technology, information resources management, information policy and services, and the study of information users.

The School has seven active research centers, of which one, the **Center for Natural Language Processing**, will be central in this research. CNLP advances the development of human-like language understanding software capabilities for government, commercial, and consumer applications. It is situated in its own lab facilities in Hinds Hall at Syracuse University. The Center for Natural Language Processing has five servers, and twenty-one computers. In addition to its own lab space and equipment, the Center has access to the meeting rooms, labs, and classroom space of the School of Information Studies. The Center also has access to technical and administrative resources within the greater University. The Center has been successful at attracting top student talent for its many Research Assistantships, including two PhD students who have won the prestigious ISI Doctoral Dissertation Proposal Award and the ProQuest Doctoral Dissertation Award presented by the American Society for Information Science and Technology.

The School's other research centers are:

- Center for Digital Commerce. Conducts research and provides strategic analyses in all areas of digital and electronic commerce.
- Center for Emerging Network Technologies. Performs hands-on testing and provide industry analysis of products and services in emerging technology markets.
- The Convergence Center. Supports research on and experimentation with media convergence to understand the future of digital media and to engage students and faculty in the process of defining and shaping that future.
- The Systems Assurance Institute, a collaboration among Engineering and Computer Science, Information Studies, the Newhouse School of Public Communications and the Maxwell School of Citizenship and Public Affairs. Advances the understanding and state-of-the-practice of systems assurance.

- The Center for Digital Literacy. Supports collaborative research and development projects related to understanding the impact of information, technology and media literacies on children and adults in today's technology-intensive society.
- The Information Institute of Syracuse (IIS) (http://iis.syr.edu/). The umbrella organization for a number of highly visible and widely successful digital education information services to improve learning and teaching in the U.S. and throughout the world.

The School of Information Studies space plan includes providing (1) a space for a community of learning, research, and education for students and faculty; (2) space that supports economic development and growth in Central New York: (3) space that supports research, development and economic growth through the School's research centers; (4) common spaces that are inviting to students and visitors; (5) space that supports communication and connections between floors to preserve the strong feelings among students, faculty, and staff of being on the IST team; (6) a building that supports state of the art technology including broadband and wireless in offices, classrooms and centers; (7) space with the flexibility to change to meet the needs of a changing networked economy, changing technology, research, and faculty and student needs; (8) classroom space that supports student access to technology and/or classroom discussions in a room such as a case management classroom; (9) sufficient conference and meeting room space for a school enriched by its faculty and staff commitment to team meetings, service, and collaborative research; and (10) space that supports a collaborative learning environment for students.

SU's Library serves the information and research needs of the academic community. The collections exceed 2.6 million volumes, 11,330 serials and periodicals, and 3.4 million microforms, located in several libraries on campus. Library services include information and reference, online database searching, access to bibliographic and other data on CD-ROM and interlibrary loan. Computing Services helps researchers, faculty and students use computing by providing personal computers, mainframe computers, data communication networks, software, training and advice. Most equipment and services are available without a direct charge.