CDI II: Modeling Information Research Behavior as a Component of Scientific Inquiry in Undergraduate Education

PI: Randall Cream, University of South Carolina

From Data to Knowledge

Our project offers the first large-scale longitudinal study of undergraduate student information-gathering as a predictor of sustainable research behavior. Our project contains three separate and connected processes: First, using a widely-available open source software plug-in for the Firefox browser, our team of researchers will deploy a campus-wide information literacy initiative to train students in information gathering as a core skill for inquiry, beginning in the required first-year composition and research courses and sustained as those students move throughout the colleges and departments of the university. Second, our research team has developed software for this browser plug-in, Zotero, which allows accurate, complete snapshots of ongoing student information research behavior to be gathered using a central server. Third, our team will aggregate and mine this data store, indexing the heterogeneous results using Latent Semantic Analysis through Singular Value Decomposition to reduce the dimensionality of the data and create useful models of student information research behavior as a correlative of educational performance, within the sciences and across the university.

Intellectual Merit: Despite a sustained interest in student information research behavior, there is a startling lack of large-scale longitudinal studies of this behavior (Williams and Rowlands, 2007). Even relatively well-gathered studies of student information research have become hopelessly outdated by the rapid transformation of the internet into a credible and sustainable resource for students to interact with journal articles, scientific papers and studies (D'Esposito and Gardner, 1999). As information research gathering has quickly shifted from using the computer to locate offline sources (databases) to using the browser to interact with those sources directly (digitized information resources), our ability to meaningfully analyze and accurately model student research behavior has fallen further and further behind. Even while our colleagues have worked to explore the utility of internet-based peer review as a component of scientific inquiry (Timmerman et al, 2003), and recent studies have suggested that upwards of "forty million Americans rely on the internet as their primary source for news and information about science" (Horrigan, 2006), these studies nevertheless remain limited to *reported* behavior, rather than verifiable representational data about information research gathering practices. Our project will produce the first-ever independent assessment of student information research collection, longitudinally tracking a student base of approximately 3500 students a year (in the first year) throughout their undergraduate career, compounded with each incoming class to total approximately 14,000 students in year four.

Broader Impacts: More broadly, the impact of our project lies in its ability to facilitate questions interrogating the relationship between student research behavior and patterns in information sources. What relationships can we identify between student performance and a semantic analysis of the research conducted by those students? What features in the data conform to cognitive choices students make in selecting sources for their inquiries? By relying on snapshots of *ongoing* research behavior, rather than self-reporting or analyses of *completed* research, we can gain an understanding of student inquiry as it takes shape, and not just in its finished form. Such an understanding is indispensible to any systematic attempt to influence undergraduate education in the sciences and beyond.

List of Participants

CDI II: Modeling Information Research Behavior as an Indicator of Scientific Inquiry in Undergraduate Education

Principal Investigator:

Cream, Randall. Center for Digital Humanities/English Department, University of South Carolina.

Co-Principal Investigators:

Buell, Duncan. Department of Computer Science and Engineering, University of South Carolina.

Eastman, Caroline. Department of Computer Science and Engineering, University of South Carolina.

Friend, Christina. First-Year English/Department of English, University of South Carolina.

Key Collaborators:

Zhou, Jun. Center for Digital Humanities, College of Arts and Sciences, University of South Carolina.

CDI II: Modeling Information Research Behavior as an Indicator of Scientific Inquiry in Undergraduate Education

PI: Randall Cream, University of South Carolina
Co-PIs: Duncan Buell, Caroline Eastman, Christy Friend, University of SC

From Data to Knowledge

We know that transformative knowledge construction is strongly associated with students learning to structure problems efficiently and productively (Shin et al, 2003). This knowledge is not acquired in isolation, though; just as we do, students build scientific inquiry by developing an ability to "sort out, evaluate, and organize their diverse views" about conflicting evidence (Linn, 2003). Researchers have concluded that students are often unable to pose meaningful problems without first attempting to construct what Linn terms a "knowledge integration perspective" facilitated by recursively considering a concept from a variety of perspectives (2003; also, Novak, 2002; Pelaez, 2002; Johnson & Johnson, 1999). While information research—drawn from original sources; published, expert-led studies; and white papers, case studies, and other documents—offers the raw materials for students to assemble this self-multiplying perspective, past studies have found that students are often unable to build the "compelling comparisons" (Davis, Bell & Linn, 2003) necessary to make sense of information research integration. These researchers have suggested the utility of an investment in "language literacy" embedded within "science literacy" (Linn, 2003), where scientific inquiry is predicated on an ability to combine, compare, and disentangle competing expert opinions (Rivard & Straw, 2000). Other studies have suggested that this inability to construct a multifaceted approach to problem building leads students to dismiss science education as irrelevant, suggesting that the impediment is not one of difficulty and abstraction but rather one of situatedness and resource/research literacy (Stocklmayer & Gilbert, 2002; Linn & Hsi, 2000; Kintsch, 1998; Songer & Linn, 1991). For science education to be successful, it must evoke students' ability to not only answer complex questions, but also the ability to ask those questions in ways that draw upon, adapt, and negotiate with the work of others (Linn, 2003; Bell & Linn, 2002; Pelaez, 2002; Zohar & Nemet, 2002). While there is extensive research that suggests the critical importance of the skill of evaluating, interpreting, and integrating the knowledge that is produced from the data of information research in scientific inquiry, there is to date no study that examines these skills *in vivo*, as an active cognitive process instead of a product of consciousness.

Our research group aims to produce the first-ever study of student information research behavior as a process of inquiry, drawn not from an analysis of surveys or the finished results of problem-solving, but rather from an investigation into the process of information searching as it occurs. We are interested in studying not just the *results* of information research but also the deliberative process of that research in order to understand the micro-processes of information juxtaposition, opinion weighting, and hypothesis generation as components of inquiry. In short, we are just as interested in the perspectives that a researcher considers useful but ultimately rejects as we are in the actual studies and reports incorporated into the finished papers and projects. Rather than mine the results of students' papers and reports, we propose to use software to maintain an accurate snapshot of student research as it is conducted, enabling us to reconstruct the cognitive dimension of research and unfold the richly complex and subtle steps of knowledge building. Our team of researchers—drawn from engineering, the sciences, the arts, and computer science and information technology—has constructed a project that is flexible enough to sustain a four-year integration into the general education curriculum and sophisticated enough to model individual and sub-group behavior within that curriculum. In the course of implementing a longitudinal study of four-years of student information research activity, we'll generate a data set that is immensely valuable to disciplines as diverse as information science, psychology, sociology, and computer science, not to mention its utility for constructing accurate models of student research behavior in science and undergraduate education. If the goal of science education, if not education itself, is to enable its practitioners to ask meaningful questions about relevant problems, and we accept that the very ability to ask those questions is constrained by our fluid navigation of information resources and data collection, then it is increasingly imperative to be able to accurately model the ways that students navigate this difficult terrain.

The innovation of our project, therefore, lies not in its desire to better understand student information research activities as a component of scientific inquiry, but rather in its ability to extend recent trends in web technology to provide the capability for accurately capturing the sub-routines of information retrieval as they occur for each student. The open-source software tool, Zotero, provides our project with a stable, widely-recognized tool for the basic function of information organization. As a plug-in to the Firefox web browser, it extends a familiar interface for information research and retrieval. Zotero works by cataloguing research brought up on web pages using the embedded metadata in page headers, but it also automatically downloads and stores linked full-text pages and attachments. The software team at the Center for Digital Humanities at the University of South Carolina, led by Jun Zhou and Randall Cream, has developed an extension to the Zotero plug-in that allows a remote sync of the data store to a unique directory on a secure server. Operating independent of user intervention, as students conduct their research and decide to add sources to their Zotero bibliographies, our extension syncs the activity automatically and securely. Over the course of semesters and years, each student's data store grows more and more complex, coming to represent a multi-faceted picture of their intellectual growth and development. The data store, anonymous but uniquely tracked, at the end of four years contains information that is almost priceless in its ability to represent knowledge-creation at the microscopic level. As such, our project deploys computational thinking to leverage bibliography-construction into information research literacy.

More pressing, however, than its ability to analyze information organization to construct a model of scientific inquiry, is our project's plan to deploy that computational thinking as a direct pedagogical device. On a curricular level, our team (led by Christy Friend and Randall Cream) will institute an implementation of Zotero that begins in first-year research and composition courses (offered within the English department) as a component of the University's information literacy requirements. In those courses, directly supervised by Friend, students will receive a flash-drive imaged to contain a portable installation of Firefox with Zotero that will launch in Windows and Macintosh environments. With approximately 90% of the first-year student population as clients, these gen-ed required courses are appropriate units to deploy our information literacy campaign. With an average population count of 22 students per class, the first-year sequence is also uniquely able to offer students a level of personal assistance uncommon in large-scale projects such as this. Through our research project, students will learn not only to use the Zotero tool to organize and retain the results of their research, but will also learn to think of the behavior of information research computationally. In learning to conduct information research as a component of problem solving, students will be informed of our research project and have the opportunity to give consent to the use of their research data. Reinforced by teaching assistants working under the training of Friend and Cream, first-year students will receive regular reminders that their data is being captured, analyzed, and modeled. In this way students will come to think of their own behavior as modelable and analyzable. As a result of this research project, students will learn to apply computational thinking to their own activities as a condition of conducting information research. By studying information research behavior, we are able not only to turn our students' attention to the important skills of information literacy as a component of problem solving, but we also are able to induce and encourage the development of computational thinking as a human behavior.

While one significant component of the innovation proposed by our research project institutes a model of information research as both a subject and an avenue for computational thinking, we are nonetheless equally excited at the consequential opportunities for creating analyses and models from our data set. Our research project will yield unprecedented access into the deliberative process of information gathering as students negotiate between conflicting opinions and perspectives to build their own research questions. With Zotero's inherent ability to effortlessly capture data from web-based research, we predict that students' research data store will greatly exceed the number of items actually used in the production of finished papers, projects and reports. With the help of Friend and Cream, we intend to train students to add much or all of the data from their deliberations to their data stores, encouraging them to use information research as a tool in inquiry, not just as a support for already-conceived opinions and evaluations. As Zotero catalogues the metadata, takes snapshots of the pages, and archives the linked attachments, each student's data store takes shape as a reflection of that student's process of inquiry. As a final item in the process, our graduate teaching assistants will encourage students to add the finished products of their research into the Zotero store, allowing us to archive an idea throughout its many iterations and variations, including its final form. The data set for our research project, then, will begin with approximately 10,000 separate but complete deliberative processes a semester (3 research assignments x 3500 students), representing each student's research work through the first-year sequence of required research and composition courses. Each process of inquiry yields important data about student decision making, information organization and retrieval, and assessment. Together, the 21,000 separate information research projects each year will serve to transform our understanding of how students think through the research problems they encounter at our university.

In addition to collecting data on each year's first-year students, we are also interested in investigating the relationship between research and inquiry longitudinally for students over the course of their undergraduate career. While we anticipate some drop-off in student involvement from semester to semester, we'll undoubtedly retain enough involvement to gather meaningful data about a time-value decomposition of research behavior across student populations. Our research team has forged links to most of the majors in arts, sciences, engineering, and education, ensuring the uniformity of information research and retrieval methodologies across campus. Assuming an attrition rate of 30% per year, our project would nonetheless produce meaningful, four-year data for approximately 1200 students a year. It would be difficult to over-emphasize the transformative possibility of this data in an area where almost all the studies are conducted by survey, and the largest observational study with longitudinal results involved 188 subjects (Williams and Rowlands, 2007). Using algorithmic analysis, including LSI through

SVD and the more demanding Genetic Analysis routine suggested by Song and Park (2008), our research project will be able to link student decisions made during research gathering to individual and group performance outcomes on a granular level. We remain especially interested in determining links between types of research behavior and demographic subgroups, including ethnicity, college major, gender, and GPA.

Even as Cream, Friend, and Zhou work to supervise and train the Zotero implementation, data collection, and data archiving aspects of the project, Buell and Eastman will work with graduate research assistants in Computer Science and Engineering to develop efficient methods for structural and semantic analyses of the Zotero data. While we propose several logarithmic methods for decompositional analyses of information research data, we anticipate that the richness and diversity of the data will suggest adaptations to existing algorithmic models, and may even dictate the creation of a new algorithmic model for analysis. Eastman and Buell will work with the graduate student as the data is collected, in real time, supervising the graduate students as they work through the large data set that accumulates.

The framework of our team—with Friend's expertise in composition, assessment and information literacy; Cream's expertise in applying computational methods to questions in the humanities; Eastman's expertise in information organization and retrieval in natural language systems; Zhou's expertise in building software tools across multiple computing platforms; and Buell's expertise in number theory and algorithmic computation—constitutes an intellectual partnership that values diversity more than disciplinarity. Uniting our disparate concerns and fields of inquiry, however, is an unyielding interest in effective undergraduate education, determined by students' problem-solving and problem-posing abilities. Together, our team understands that our true education agenda is overcoming the challenge of viewing education as an object-oriented existence, where students come to see the world as composed of discrete fields and disconnected objects and problems. Each educational activity, in this view, is a separate and self-contained struggle. Our research goals, however, are to develop cross-disciplinary tools that affect the ways students understand themselves in relation to the problems posed in education. Thinking about themselves computationally—understanding their actions as recursive process that iterate across common objects—is a significant step towards that end.