Information Artifact Enhancement
Collaboration based on
Access Control, Annotation, Logging, and Analysis

Introduction

Theoretical approaches to building human centered systems and integrated information environments have met with limited success. While progress in constrained areas of each problem area has been significant (speech recognition, content analysis, text summarization, agent interaction, etc.), the success has been dwarfed by approaches that were far less organized. The WWW was neither intuitive, nor agent enabled, yet in a few short years it captured the attention of masses of people and became a phenomenon. Currently, RSS, designed initially as a metadata site summary structure, is radically changing how we approach and envision the web. We propose to take a macroscopic approach to assessing the potential of four key technologies to enabling better collaboration based on information artifacts. Put more simply, like the web and rss, it is our intent to bundle existing technologies that appear underutilized and package them so as to assess if they might become the next generation to technologies to fulfill important needs. If the systems do not reduce human workload or provide significant functional advantages, they wither and die. The analysis of the success of email versus calendaring systems (cite) provides one example. The success of the web versus gopher or ftp (cite??) provides another example. This project proposes to deploy on the web and assess a collaboration system based on information artifacts that provides open hooks with demonstration services to four key capabilities – logging, annotation, analysis, and access control. The infrastructure, key capabilities, and functional demonstration elements are outlined below. Following the description of the system, we describe how we will assess the potential of the efforts to shape the progress of human centered computing and information integration in a variety of domains.

Various visions have been proposed for the next generation web under three main rubrics -- the “semantic web”, web2.0 and web 3.0. Generally, they exploit several of five key technologies – a manipulable DOM, a DOM aware programming language, asynchronous remote method invocation, application servers and metadata. The webservices vision is focused on dynamically configurable services located using metadata and linked by application servers. The AJAX vision is focused on manipulable DOMs linked to backend services via DOM aware programming languages and remote method invocation. We propose to assess the impact of these key technologies as means to provide end user services for collaboration based on a system for working with the
information artifacts of an enterprise where the user experience is enhanced via unobtrusive subsystems for access control, annotation, logging and analysis. It leverages the strengths of the Information Science faculty in the School of Information Sciences at the University of Pittsburgh. While the specific foci of the efforts to be undertaken during the grant are designed to leverage these local resources for purposes of administrative efficiency and cost containment, the proposal does include a broader architectural vision of the system.

**The Proposed System**

Commercial and research collaboration systems exist. Various applications and research projects have incorporated one or more of the core subsystems. The key technologies all exist and are in various stages of refinement. However, we do not know of any systems that have used these key technologies make these core capabilities easily available to developers as building blocks or to end users via unobtrusive services. Further, we do not know of any systematic efforts to assess the end user experience in terms of it being human centered or information integrating. The system we propose is a generic collaboration framework that will be diverse enough to support scientific research, education, document management in areas a diverse as the development of standards, legislation, legal documents, etc. We have developed traditional client server collaboration systems over the years. CASCADE (Computer Augmented Support for Collaborative Authoring and Document Editing) was used for standards development and instruction and was the basis of research to study other collaboration activities – software inspection (cite Sapsomboon) and collaborative authoring (cite Spring, Vathanophas). More recently, we have built systems to generate integrative web systems to support research on web based medical interventions (cite Grady). We have also developed embedded information systems that provide just in time information and role associate information. Currently, we are working on fine grained web based document commenting systems with access controls. We have explored the potential of designing similar systems to support scientific collaboration related to volumetric scientific data.

At its core, the envisioned system is a web service based information artifact management system. The core of the system includes artifact storage and retrieval, naming, locking and version control. It is generalized for various artifact types and includes both content items and container items. The core subsystems are:

- **The Access Control** subsystem includes both user access rights and artifact protections in a multi layered system that makes fine grained artifact distinctions and uses groups for ease of administration.

- **The Annotation** subsystem provides for private and shared annotations in both graphical and textual form. It has facilities for metadata classification of comments as well as annotation analysis.

- **The Logging** subsystem tracks events – i.e. user activity, artifact status, etc.
• The **Analysis** subsystem includes reports on user activity, artifact activity, administration activity, etc.

The core subsystems will be constructed to support extensible web service interfaces providing many of the features that have previously been developed for CASCADE. These would include:

• **Communication** services include both one to one (mail, chat SMS) and group communications (white boards, structured meetings, ballots, etc.)

• **Navigation** services includes both formal and adhoc navigational systems _and_ these apply system wide, around clusters and within artifacts.

• **Administrative** services involves everything from adding and removing users and groups to setting up project semantics and functionality

• **Reporting** services include usage reports, activity reports

• **Artifact** services will include translation, summarization

• **Visualization** services

The infrastructure supporting these subsystems will include:

• Data Import and Export

• Storage

• Encryption and Authentication

• Backups and Version control

Figure 1 shows a picture of the layered services, subsystems, and core.

![Layered Architecture Model](image-url)
Special Focus End User Services

There will be seven functional units of the system that will be singled out for testing during the conduct of the grant they are:

- Authentication and Access Control Services
- Annotation Services
- Metadata Generation Services
- Communication Services -- Meeting Management
- The User-Artifact Activity Analysis Unit
- The Document Artifact Structuring Unit

The rationale for examination of these units is briefly described below. We assume that people will continue to develop documents and other information artifacts for exposure over the web and that the demand for new services and capabilities will continue to evolve. From a high architectural view, web systems and services fall into three increasingly complex categories – information, interaction and transaction. While Berners-Lee envisioned tools that would operate at all three levels, the first generation web was dominated by information services. Many commercial sites have evolved to transactional services, but the vast majority of the web operates at information and interaction levels. We plan a series of studies and experiments to determine which of the following services are likely to be widely adopted by a broad user community. They include advanced information services, interaction and transaction services we believe are underrepresented on most non-business sites. It is our contention that the services must be available as open source tools that are intuitive, scalable, extensible and robust.

The Authentication and Access Control Services

Authentication is an increasingly important aspect of web life. The basic technology for distributed authentication and access control has existed for a number of years. Despite the importance of authentication and access control and the availability of the technology, its use is limited. Some of the reasons for a lack of widespread adoption include the complexity of management in complex and evolving environments, the lack of self-organizing approaches, the isolation and proliferation of individual sites, and concerns for inadequate security in a web based environment. It is our intent to develop an authentication module that provides system administrators, and ultimately users with the following functionality:

- The module can be configured to require manual setup, coordination with an existing LDAP server, invitational self registration based on emailed tokens, or self registration based on recognized certificates or FOAF certificates.
- The module will allow system administrators to provide their own limited certificate distribution capability.
- The module will allow inherited group rights to individuals and inherited protections and access controls for artifacts.
- The module can be configured to do manual or automatic clean up of accounts in a variety of ways, parallel to the setup options.
Described in the evaluation section are a series of usability studies that will be conducted on the various component pieces during years one and two followed by deployment evaluations at a number of sites during year 3.

The Annotation Services
Annotation is a step both toward collaboration and toward personalization. Effective annotation of documents has many uses. Unfortunately, the lack of access controls on the web and the nature of html and the http protocol have stymied efforts such as WEBDAV. Earlier hypertext efforts, the emergence of XML, AJAX and ATOM all signal the chance to change this. The goal of this effort is to make the kind of annotation capabilities available in Microsoft Word available as shared tools in controlled spaces operating over multiple artifact forms. It is our intent to develop an annotation module that provides the following capabilities:

- Provision for private, semi-private, and public annotation forms.
- Provision for graphical (e.g. highlighting with *’s or editorial symbols) and textual attachments.
- Provision for various attachment forms (point, range, area, volume, and time based anchors) appropriate to various artifact forms (text, audio, structured text, images, volumes, and video).
- Provision for timestamps, author info, metadata, and access info.
- Provision for collective analysis.

Described in the evaluation section are a series of usability studies that will be conducted on the various component pieces during years one and two followed by deployment evaluations at a number of sites during year 3.

One special application of annotation might be to tie it to an intelligent query system. We have found with CASCADE used in an instructional mode that a system that allows student to pose questions in context is very efficient for students, and answers attached by the instructor to the question are very efficient and also informative to other students. The goal of this module is to extend the capability in several ways. It is our intent to develop an intelligent query service module that provides the following capabilities:

- Users may pose a general question or add a question annotation in the system and the system will take any contextual information (artifact related location, time or term, student profile) and use it to query a database of prior questions and answers. It will immediately provide what it believes are relevant archived answers or inform the user that the message is being transferred to an expert.
- Experts will receive communications from users along with any contextual information and any archived answers considered relevant. The answer provided will be routed back to the user, inserted in the archive, and attached to any situated posting of the question.
- Periodic reports will be made to the owner of artifacts that produced patters of questions, period of time that produced similar sets of questions, in situ Q&A
objects that were visited and studied and these will be presented in a format and form that induces owner modification of the artifact.

Described in the evaluation section are a series of usability studies that will be conducted on the various component pieces during years one and two followed by deployment evaluations at a number of sites during year 3.

**Metadata Generation Services**

The development of the semantic web vision that seeks to improve the precision and recall of web searches is based on the development of five large systems – 1. a series of ontologies that will define the metadata of collections, 2. the metadata consistent with those ontologies that is accessible for collection and analysis, 3. a set of high level reference ontologies that allow existing ontologies to be mapped, 4. a set of ontological statements that allow for inferencing and deduction, and 5. an infrastructure for the organization, dissemination, and processing of this information. Progress is being made in all these areas. It is our conviction that unless we solve subproblem 2 – metadata generation, solving all the other problems is for naught. Further, metadata generation will need to be accomplished at the grass roots level unless it can be fully automated – which we do not think is possible. Therefore, what is required is a tool that demonstrates the benefits of such registration and that does so at a cost that is acceptable to those who have to use it. It is our intent to develop a metadata registration module that provides the following capabilities:

- Intelligently scan existing document collections to harvest known metadata (e.g. Word inserted information, image tag data, etc.) and provide easily analyzed metadata related to structural and textual data. The tool will be primed by a system administrator who will be able to statically or algorithmically provide additional data (e.g. the organizational publisher on the information on the site is XXX or the author of documents may be determined by a mapping table that associates path location with individuals, i.e. /~mbspring/* documents are authored by Michael Spring.)
- Present reports of efforts to the appropriate individual for review highlighting questionable or missing metadata and associating it with a visible document to aid the metadata review/completion process in batch mode.
- Store the results of such reports in optimal formats for searching locally and export with relevant ontologies to repositories.
- Provide automated additional metadata for information artifacts created as a part of the system.

Described in the evaluation section are a series of usability studies that will be conducted on the various component pieces during years one and two followed by deployment evaluations at a number of sites during year 3.
Communications Services – A Structured Example

Some activities benefit from coordinated synchronous user activity. Historically, the stateless pull technology of the web was ill suited to support such activity. The emergence of applets with the ability to connect to server processes and more recent AJAX-like initiatives make support of coordinated synchronous activity possible. It is now possible to alert connected users to changing state without explicit actions by them. Further, centralized logic and business processing rules can be shared across multiple sessions. It is our intent to develop a meeting management module that provides the following capabilities:

- Provide information about aggregate access patterns that allows for optimal timing of meetings while minimally divulging private information.
- Provide facility for approving access to user activity data so as provide other meeting participants with social periphery information.
- Provide an array of selectable communications mechanisms both as pure communication channels and as communication through artifacts.
- Provide for selectable levels of meeting management from formal (e.g. Roberts Rules of Order) to informal (e.g. tokenless whiteboard discussions).

Described in the evaluation section are a series of usability studies that will be conducted on the various component pieces during years one and two followed by deployment evaluations at a number of sites during year 3.

The Document Artifact Structuring Unit

Web systems have historically been very tolerant of document structure. Unknown tags are simply ignored and common coding errors – lack on end tags, are covered up. With the emergence of new XML compliant document types, both well-formed and valid documents are increasingly required. Further, optimal use of new technologies such as AJAX require extensive use of complex features such as IDs. More importantly, these capabilities have to ultimately function over dynamic end user defined documents. Particularly when these documents are used in collaborative efforts, users will benefit from a simplified document structuring aid. It is our intent to develop a document structuring module that provides the following capabilities:

- Development of an extensible instantiation of a schema using simple forms and questions.
- Optimal structuring of the instantiation into appropriate size units to allow granular locking of pieces of the document.
- Development of pruning and grafting modules with validation checks
- Integration of ID and IDREF capabilities to support AJAX like technologies that make use of the DOM at levels 1 and 2.

Described in the evaluation section are a series of usability studies that will be conducted on the various component pieces during years one and two followed by deployment evaluations at a number of sites during year 3.
The User-Artifact Activity Analysis Unit

In a classroom or a meeting room, members of a group, be it hierarchical or democratic, have the ability to assess how ready people are to participate and how in touch they are with the matter at hand. In distributed systems much of that goes away. Perhaps of more interest is the ability to intimately track user interaction with an expository artifact. Research on eye tracking provides an extreme level of detail on interaction. Less detailed information is provided by time spent by the user, number of accesses, scrolling activity, annotation activity, etc. It is our intent to develop a user-artifact activity analysis module that provides the following capabilities:

- User activity with a given artifact will be logged and made available as a social periphery analysis tool in meetings allowing group members to assess the level of engagement with the artifact, degree of understanding or commitment, and willingness to participate.
- User activity with a given artifact will be analyzed with the intent of providing the owner of the artifact with whatever information might be of use in modifying or correcting the document.

Described in the evaluation section are a series of usability studies that will be conducted on the various component pieces during years one and two followed by deployment evaluations at a number of sites during year 3.

Evaluation of the System and the Special Focus Modules

Description of the experiments … I am thinking about 3-5 experiments that would be based on the infrastructure. Some examples might be things like:

1. Research Interest Group Wiki: This experiment will use the logging, communication, and artifact management modules. It looks to assess optimal levels of editorial controls in a small group. Faculty and Graduate Students in the School of Information Sciences are members of research interest groups. Given the schedules and interests of the involved participants, they conduct some amount of the business asynchronously. Current wiki technology tends to be crudely controlled and operationalized. This experiment will log data about access patterns and activity in wikis that have various levels of granularity and access controls to assess the optimal mixture of capability, flexibility, and control in small groups of peers.

2. Building Management: This experiment will use the logging and analysis modules. It will feed data from a sensor net and sanitize it for use by a building management system that will look to determine patterns of activity in a building and make use of calendaring information to make recommendations for changes in scheduling and use of the facility. (ugh – I thought it was going to be better.)
3. Multi-level Annotation for Education: This experiment will make use of the artifact management and artifact annotation modules. It looks to explore how cross educational level collaboration can take place. Specifically, a team of students in an undergraduate course on code design and testing will be paired with a graduate course on systems analysis and design and a high school course on a programming language. The graduate course will develop specifications for a software module that will be coded by the high school students. The high school student code will be examined by the undergraduate students conducting code inspections. The results of the code inspection will be reviewed and amended by the graduate students and fed back to the high school students. The experiment will vary the nature of the interaction – synchronous, asynchronous, and mixed. It will also vary the code complexity and….

The evaluation of the system will be based on unit testing in house and then on propagation to the web and assessment of whether people adopt or use the technologies

**Plan**

The first part of year one will focus exclusively on the translation of CASCADE infrastructure to a web services architecture. This work will continue through the rest of year one at a reduced pace while simple forms of the special focus functional unit are brought on line and put through usability testing. The modules will be refined and further integrated into the infrastructure in year two. In year three, the complete systems will be disseminated and tested in at least 10 locations. Logging mechanisms will be developed to gather and report back usage choices and usage patterns to allow for an assessment of whether the system as a whole worked for users and more particularly, which of the special focus functional units offer promise as widely adoptable technologies.
Appendix 1:

(Rework this to make it consistent with the text and flesh it out.)

The CASCADE system, based on an early feature analysis of collaboration systems (http://www2.sis.pitt.edu/~spring/cascade/feature_analysis.html) provided the basis for the functional units planned under this project. While the specific nature of the services will emerge out of the design, some of the services that we anticipate will be developed include:

1) Infrastructure
   - LDAP environment (naming, addressing, discovery, composition)
   - Secured communications
   - Distributed authorization and access

2) Management
   - Creating users and groups and assigning access rights
   - Developing spaces and artifacts within those spaces
   - Work management

3) Basic Services
   - Artifact Services
     - Creating specific pieces of work
     - Writing, editing, adding
     - Editing tools – fog, usage, audience, etc
     - Authoring tools – schemas, schema to instances, instances
     - Specialized Authoring tools – charts, graphics, animation, equations, sequences
   - Navigation
   - Annotation, commenting, questioning tools
   - Comment analysis tools
   - Activity review
   - Communication and meeting tools
     - Directly through tools
     - Indirectly through votes, comments, etc.
   - Monitoring and feedback tools

4) Support Services
   - Data stream capture and cataloging
   - Data Management
   - Data access and locking
   - Market management services (accounting, system operation, etc.)
   - Delivery and conversion tools
   - Project Management
   - Adaptive profile service
   - Activity monitoring