Project Prism activities described in the remainder of this section are:

- *Digital Object Architectures* – We are continuing to use the Fedora digital object architecture for experimentation with complex digital content and its localization.
- *Digital Preservation* – We are investigating policies and mechanisms for preserving digital materials in the research library context.
- *Human-centered Research* – We are working to understand and enhance the human effects of digital libraries and distributed information systems.
- *Interoperability Architecture* –Through the Open Archives Initiative we are developing and disseminating low barrier digital library interoperability solutions.
- *Policy Enforcement* – We are investigating techniques for enforcing policies for preservation and security on digital objects.
- *Automated Digital Libraries* – We are using focused web crawling to automatically aggregate open access web resources into semantically meaningful collections.
- *NSDL Testbed* – We are exploiting Cornell's key role in the development of an infrastructure for the NSDL (National Science Digital Library) by using it as the context for testing and deploying Prism research areas.

1 Digital Object Architectures

Over the past year we continued to undertake research and development on our FEDORA digital object architecture. As defined in earlier reports, the key feature of FEDORA is its ability to encapsulate and add value for complex digital content. Our work over the past year on FEDORA is divided into two areas: research in context-sensitive behaviors and collaboration with the University of Virginia to develop and disseminate an open-source FEDORA application for general use in creating digital object repositories.

1.1. Context-sensitive behaviors

The motivation for our work in this area is as follows. Assume a Digital Object that contains lecture content: a video of the lecture, a set of *gifs* shown at the lecture, and XMIL synchronization data. It is desirable that access to this lecture digital object via a French site consists of the translation of the text on the gif images from English to French, and a French translation of the lecturer's speech. Similarly, access of this lecture from a Gallaudet University site should consist of a transcription of the audio from the lecture is provided. Our work over the past year has demonstrated that we can add value to digital information by providing an "appropriate" or personalized experience of content.

Our approach has been to split the creation, storage and manipulation of content from the definition and implementation of behaviors associated with content. The "experience" or "presentation" or "rendering" of the content takes place closer to the user in our modified architecture, so the experience can be tailored to particular communities of users. The two key components to our approach are 1) exposing structural metadata associated with Digital Objects – metadata about relationships among content pieces in a Digital Object and 2) information intermediaries called Context Brokers that match structural characteristics of Digital Objects with software mechanisms that produce behaviors. Note that the Context Broker manages the *interaction* between content and behavior mechanisms: the Context Broker might control neither the content nor the mechanisms.

We report on the outcome of this work in the results section of this report.

1.2. University of Virginia Mellon Collaboration

In the previous reporting year we began to collaborate with the Advanced Digital Library Research and Development Group at the University of Virginia. The group at Virginia had been looking for a solution to their complex digital content needs in the library and as part of their NEH and Mellon funded Institute for Advanced Technology in the Humanities. After an unsuccessful search among vendors, they began to experiment with our FEDORA architecture and produced some impressive prototype results. Initial collaboration in the process of developing this prototype led to the conclusion that the group at Virginia provided the appropriate context for refinement and technology transfer of our Prism FEDORA results. A substantial grant from the Andrew W. Mellon Foundation provides support for the pure development and deployment aspects of the collaboration, while NSF funding provides support for architectural research and specification. The collaboration also includes a number of deployment institutions including:

- New York University
- Tufts University
- University of Indiana
- Kings College (UK)
- Oxford University (UK)
- Library of Congress

2 Digital Preservation

Digital preservation research in Prism is joint between the Cornell University Library (CUL) and the computer science department. This has afforded a unique combination of technical and policy skills in approaching the problems of digital preservation. The joint team leads investigations of policies and mechanisms for digital preservation.

The activities in this area, with details provided in the Findings section of this report are:

- *Preservation Risk Management for Web Resources* – The team has been carrying out experiments using the Mercator web crawler to automatically evaluate risks to Web resources and has developed an approach and a Web page profile for analyzing Web resources.
- *Digital Information Longevity Study* - The longevity study is looking at technological change, as well as organizational, administrative, and economic issues that may have had affected the persistence of the resources.
- *Project Harvest* – In work co-funded by the Andrew W. Mellon Foundation the preservation team developed a Subject-Based Digital Archives (SBDA) model.

- *CU Library Central Depository Model* - The Prism team led a library-wide effort to develop recommendations for establishing a central depository for preserving Cornell's digital image collections.
- *Information Theoretic Preservation Model* – A paper presented at ECDL 2001 reported the results of work defining a digital preservation model that is information centric rather than bit centric.

In addition to and in support of the work reported in the section, members of the research team participated in a number of international activities related to the research work occurring within the Prism context:

- Anne Kenney serves on the EU/NSF working group on Digital Archiving and Preservation. This is one of seven working groups being established by DELOS, the EU-funded Network of Excellence for Digital Libraries.
- Anne Kenney serves on the standing committee for Cuban Libraries and Archives that is jointly sponsored by the American Council of Learned Societies, the Social Science Research Council (SSRC), and Academy of Sciences for Cuba.
- Anne Kenney serves on the RLG/OCLC Working Group on Digital Preservation: Defining a Sustainable Digital Archive. This international committee has produced a draft report for public comment in August 2001 entitled: *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources* and is working on the final version of the report.
- Anne Kenney is chairing a joint working group of the Council on Library and Information Resources, the Association of Research Libraries, the Oberlin College Group, and the University Libraries Group to study the state of preservation programs in American college and research libraries. A key component of that study is to assess digital preservation policies, programs, and institutional readiness.
- Nancy McGovern is working with the Certification working group of The Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS), part of the ISO Archiving Workshop Series. AWIICS is coordinating a series of follow-on activities of the Open Archival Information System (OAIS) development initiative.

3 Human-Centered Research

The HCI Group continues to conduct research on situated information seeking in order to examine relationships among task, purpose of user, perspective of user, presentation of content, and information categorizations. The purpose of these studies is to identify behaviors that can enable the creation of navigable multimedia representations of information and can offer promising directions for the development of innovative information categorizations and descriptions, metadata and consequently, innovative search and retrieval tools. This work is divided into three areas as described below.

3.1. Search Behavior

Research this past year has focused on addressing the kinds of search strategies employed under varying conditions of expertise, resource availability, and user characteristics. Descriptive data of users' search strategies subjected to Multidimensional scaling indicated distinct clusters that visually appear very different under the varying conditions and between gender. Search terms and process terms were quantified in order to compliment some of these more qualitative data. These data also indicated differences among the conditions and between gender. While these data helped us to better understand the search process, it was preliminary only. Since this pilot we designed a more systematic study to investigate specifically the search terms employed by users who have identified themselves as expert or novice on several different topics. In addition, we have tried to mimic so of the typical feedback that users might receive under the various search conditions.

3.2. Tracking Behavior Online

We have continued to log all Internet and email use for a fourth consecutive semester from participating students in an upper level communication course as well as students enrolled in an engineering course as part of a collaborative effort between Cornell and Syracuse Universities. From these log files we have examined the development of social networks among the students and the effects of prestige on social navigation.

3.3. Context Aware Computing

Our work this year has focused on the design and evaluation of two prototypes we developed using GPS for the outside application and infrared beacons for the indoor installation at the Johnson Museum on Cornell's Campus: CampusAware and MUSE. Both prototypes are described in the results section of this report.

4 Interoperability Architectures

We continued over the past year to both support and develop the Open Archives Protocol for Metadata Harvesting (OAI-PMH). This work continues to be funded through Project Prism with additional administrative funding from the Digital Library Federation and the Coalition for Networked Information. As in the previous year, Cornell acts as the executive location for the OAI, with participation from a range of other international parties and organizations.

Our 2000-2001 report described the development of OAI-PMH as a mechanism for interoperability amongst ePrint repositories to a general interoperability mechanism. This led to the first public release of the OAI-PMH (version 1.0) in early 2001. These activities are fully described in a paper presented at JCDL 2001 in summer 2001. Events during 2001-2002 indicate a broad level of acceptance of the protocol as both a production tool and a vehicle for research. These include:

- Funding from the Andrew W. Mellon funding to seven “service providers” – that harvest metadata from OAI-PMH data providers and build value-added services with it.
- Registration of over 50 data providers at the OAI central registration service (developed and supported at Cornell). We believe that these fifty represent less than half of actual implementers.
- Funding by the European Community of a parallel effort, known as the Open Archives Forum (OAF). The goal of OAF is to encourage and support the development of “open archive communities” in the European context.
- An organized effort within the museum community, led by CIMI, to build federated museum services using OAI-PMH.
- A substantial amount of OAI activity at JCDL 2001 including papers, workshops and tutorials.
- A JISC funding call with OAI as a central activity.

Our goal since the initial release of the protocol has been to use this year for experimentation. This means both maintaining the stability of the protocol over the year and working towards a subsequent “standard” release (version 2) of OAI-PMH. We can report success in both areas.

Our careful work in developing and vetting through careful alpha testing of the protocol early in 2001 proved its worth in the lack of serious problems and resulting stability of the version 1 protocol. During the year, we only had to issue one follow release, 1.1, that was necessitated by a change by the W3C in the XML schema specification, upon which OAI depends.

The outcomes of our work on version 2 of the protocol are described in the results section of this report.

5 Policy Expression and Enforcement

Prism continues to focus on the challenges of security policies, including the definition, formal declaration, and enforcement of policy . The issue of *policy specification* (definition and formal declaration) starts from the point where humans intellectually define acceptable and unacceptable scenarios for the use of digital library content. The next challenge is the formalization of those definitions into policy declarations and the ultimate refinement of those declarations into specifications that can be understood by automated enforcement mechanisms. The issues of policy specification are intimately tied to the challenges of *policy enforcement.* Here we look at different technical approaches for ensuring that applications are secure and able to prevent a breach of policy when users access them.

5.1. Policy Enforcement for Digital Objects:

We continued our work in specifying and enforcing fine-grained policies that pertain to individual digital objects, or to sets of similar digital objects. Our goal has been to demonstrate a policy specification language and enforcement mechanism that supports policies that are very expressive, and highly customized to particular kinds of objects. Instead of specifying policies that are rather general, like whether a particular user can read/write/delete a particular object, our goal has been to support policies such as:

- ***Policy 1:*** For a course object, GUESTS may view the course syllabus and the introductory slides of the first lecture, but may not view the lecture video, or any of the other slides.
- ***Policy 2:*** For a course object, STUDENTS may not view the video of the second lecture unless they have submitted their assignments pertaining to the first lecture.

5.2. Stateful Policy Enforcement for Multi-Tiered Applications:

The goal of Inline Reference Monitoring (IRM) is to detect and prevent undesirable state transitions in a target program. IRM has proven successful in its ability secure untrusted mobile code (specifying and preventing bad things the code can do to a receiving host). We have also shown that IRM is useful for enforcing application-specific and object-specific access control policies (see Section 5.1). Our recent experiments in using IRM for policy enforcement in digital library applications has unearthed the significance of application-specific "state variables" in the policy enforcement process.
In IRM, policies specify conditions under which executions will be prevented. Many of these conditions are expressed by reference to the *current values* of application-specific variables, for example a condition might be: variable *user_id* is "susan" and variable *has-paid* is "true" and variable *user_role* is "student." While IRM is a secure way to modify bytecode to insert policy checks into programs, it has not addressed the challenges inherent in managing and using application state variables, especially when such variables, or their current values, are not directly available in the IRM target (the module that runs the code relevant to a specific policy). We define two new dimensions to the IRM problem:

- *availability* of all relevant application state variables necessary to enforce a policy
- *trust* in the security of the current values of application state variables

Essentially, policy enforcement is only as good as the security of the information that a policy uses in making its enforcement decisions. If a policy is making decisions to prevent executions based on the current values of variables in an application domain, the policy is dependent on whether it can actually obtain those current values, and on whether those current values can be trusted.

The problem of obtaining and trusting the values of application state variables is complicated in highly modular applications, multi-tiered applications, or collaborating applications. The more dispersed or distributed an application is, the more difficult it may be to locate variables required by a policy, and to ensure that the current values of these variables can be trusted. During the previous year we have been exploring the following questions. What programs are responsible for updating variables? How do different applications or different application domains exchange these variables? How many applications or domains have access to these variables? How do we ensure that all necessary variables are "in scope" for an IRM-modified target?

5.3. A Logic for Policy Specification

The goal of this work is to allow people who have not been trained in formal methods to state policies precisely and reason about them in a provably correct manner. (By a policy, we mean a set of conditions under which an action such as reading a file is permitted or forbidden.) To achieve this goal, we are writing a software package whose back-end is based on logic, but whose interface is appropriate for non-logicians. More specifically, the user enters policies and facts about his/her environment (e.g. Alice is a librarian) by filling-in blanks in English sentences that the software provides. The input is translated into a language that is based on a fragment of first-order logic. Then standard model checking techniques are applied to the formal statements to answer questions about the policies. For example, the model checker would determine if a particular action, such as Alice downloading a budget report, is permitted or forbidden.

6 Automated Digital Libraries – Collection Building

During the past year, we began two initiatives in the area of "automated digital libraries". The goal of this work is to understand how many of the classical library activities can be automated in the digital and web environment. The motivation is to reduce the cost factor that is associated with human effort. Our primary technique for doing this is "focused web crawling" and our tool is the Mercator web crawler. The previous section on preservation related activity described the application of this tool and technique for risk management. This section describes our work on automatic collection building.

A section of the original proposal suggested work automatically generated collections. At the time of writing of the proposal, the techniques for doing this were envisaged as predicates over the contents of special "digital library repositories".

Our increased involvement and cross fertilization with the NSDL project (described in Section 7) has led to an interesting reformulation of this original plan. Instead of controlled "digital library repositories", NSDL works in the context of the web, working both with content in web servers maintained by cooperating parties, and by organizations who do not actively participate in NSDL. In this context, our focus has shifted to automatic subject based collection gathering in web space using focused web crawling. We believe that such non-human intensive methods are necessary if large-scale digital libraries such as the NSDL are to succeed. This research is being undertaken by Donna Bergmark in the CS department.

The question being explored is: "suppose you have a powerful crawler running on a fast computer with a high-bandwidth connection to the Internet backbone; how far can you get in automated collection building in a digital library?"

Starting in August, 2001, we had access to Mercator, a fast web crawler located in Palo Alto, CA. It was developed at the Systems Research Center under Digital, and received further development under Compaq. Written in Java, Mercator is extensible and therefore a perfect vehicle for research into bulk collection building.

Putting together online collections of topic-specific educational material is usually done by hand. However, this does not scale at the same rate as the web itself, which is growing exponentially. The goal we set for the Fall 2001 semester was to generate, automatically, a number of collections on topics in science and mathematics. We define a collection as a set of HTML, PDF, and Postscript page on a given topic.

The preliminary results of this work are described in the results section of this report.

7 NSDL Testbed

In third quarter 2001 Cornell received NSF funding as part of the NSDL (National Science, Mathematics, Engineering and Technology Education Digital Library) core integration (CI) effort. NSDL, and particularity CI provides a perfect testbed for the exercise and testing of many of the interoperability concepts that are a subject of Prism research.

One tangible result has been the development of the NSDL core architecture, that builds on the interoperability framework described by OAI-PMH. Briefly, the OAI core architecture consists of a metadata repository, with contents that are the result of OAI-PMH harvesting and automatic generation of metadata from open access web pages. The contents of the metadata harvesting will also be available through OAI-PMH to a, hopefully, developing suite of enhanced service providers.

One additional result of this collaborative activity is the automatic collection building described in Section 6.