Project Prism
Annual Report 2000-2001

This report documents progress and future plans on Project Prism at Cornell University, National Science Foundation Contract No. IIS-9817416. Research areas in this report are organized in the same manner as our project web site at http://www.prism.cornell.edu. For more details, description of other affiliated activities, and up-to-date information refer to that page. These areas are:

- *Digital Object Architectures* – The Fedora digital object architecture permits the packaging of complex digital content and accommodates extensible disseminations from that content. During the past year we conducted extensive interoperability testing and initiated explorations to support dynamic, context-sensitive binding of digital object behaviors in a Web context.
- *Digital Preservation* – We are investigating policies and mechanisms for preserving digital materials in the research library context. During the past year we started work on a number of long-term collaborative studies related to digital preservation.
- *Human-centered Research* – We are working to understand and enhance the human effects of digital libraries and distributed information systems. Over the past year we conducted a number of experiments in search and online behavior based on data collected in courses at Cornell University.
- *Interoperability Architecture* –The Open Archives Initiative develops and disseminates low barrier interoperability solutions. During the past year we developed and published the Open Archives Initiative Protocol for Metadata Harvesting.
- *Policy Enforcement* – We are investigating techniques for enforcing policies for preservation and security on digital objects. During the past year we made substantial progress on defining logics, languages, and mechanisms for policy enforcement..
- *Web Preservation* – We are investigating mechanisms for preserving resources on the open Web. During the past year we collaborated with the Library of Congress to develop a strategy and coordinate a prototype of Web preservation.

1 Digital Object Architectures

We continued to refine and extend our Fedora digital object model, which provides a basis for our experiments with complex digital content. Work in this area is led by Naomi Dushay, with assistance from Sandy Payette, Sugata Mukopadhyay, Jill Newman. A detailed Web page on Fedora is available at http://www.cs.cornell.edu/cdlrg/FEDORA.html.

The following refinements to the Fedora Digital Object model were successfully developed and tested in a continuing interoperability experiment with The Corporation for Networked Research Initiatives (CNRI):

- more flexibility in Digital Object content: multiple "bootstrap disseminators" are now allowed on a Digital Object.
- improved security the for loading of remote executables
- explicit parameter domains can now be declared for Digital Object behavior requests, which can improve both clarity and interoperability
- improved the clarity of the protocol with terminology changes, better consistency of terms throughout the protocol, and better use of object-oriented interfaces and inheritance.
- improved engineering of the software, including:
    - better organization of software and sample content for easier conceptual comprehension
    - runtime configuration of repository and clients, rather than compile time
    - better error handling
    - improved documentation, including release notes and a primer on the most difficult Fedora concepts.

To further test and challenge our digital object model, we collaborated with Cornell Computer Science Ph.D. student Sugata Mukhopadhyay, whose research focuses on automatic processing of multi-media content. Sugata's work captures lectures on simultaneous videos and then automatically creates an edited presentation video synchronized with gifs from the originally projected presentation. His original approach was to store the content in an ad-hoc fashion and provide access to it via a java-script program accessed via a Web browser.

Fedora provided a structured way to store these complex multimedia digital objects, and also provided a means to create behaviors allowing for a synchronized multimedia presentation of the lecture. In fact, we wrote three different lecture clients for Fedora. The raw content of a lecture digital object contains the presented slides as gifs, a URL for the edited video (allowing us to skirt the issue of streaming video data directly via Fedora) and a SMIL file containing synchronization information encoded in XML. The "Heavy" client does processing on the client side: the client parses the SMIL data, and all the slide gif images are sent over to the client at startup. The "Lite" client does the least processing on the client side: the SMIL is parsed at the repository by the extensible digital object behaviors loaded as needed at run-time, and the slide gif images are obtained as needed. The "Hybrid" client parses the SMIL data, but obtains the slide gif images on demand. These clients also allowed us to demonstrate policy enforcement techniques: policy granularity could be adjusted according to the defined behaviors for each client's approach.

We also investigated representing Dienst documents via Fedora. Just as there are different ways to achieve the "lecture experience" via Fedora, there are many ways to represent Dienst behavior, or indeed, many behaviors, within Fedora. We did some thinking on these issues and wrote a paper on them (see "Modeling Decisions for Digital Content" in publications section).

We are now working on a significant change to Fedora: we want to be able to bind behaviors to Digital Objects dynamically (as well as statically). This lets us split the creation and manipulation of content from the definition and implementation of behaviors associated with content. It also allows Digital Object behaviors to be automatically updated. Furthermore, we intend to allow context-sensitive binding of behaviors. Motivation: suppose you have a Digital Object that contains lecture content, as above: a bunch of gifs, a video, and SMIL synchronization data. The lecture was originally given in English, but if you access the lecture via a French site, then software automatically translates the text on the gif images from English to French, and also provides a French translation of the lecturer's speech.

Dynamically bound behaviors have been fully designed, and implementation is just beginning.

We will accomplish dynamic behavior binding by implementing an abstraction known as "Structoids" – a definition of the relationships among the content pieces in the Digital Object. Structoids can be viewed as "behavior potential" – the Structoids in a Digital Object are what allows it to bind to specific behaviors. Structoids can be thought of as recognizable patterns in objects that have to do with the relationships among data. So Digital Objects can be defined a) by their content b) by their behaviors c) by the relationships among the contained data.

Our future work on Fedora will make extensive use of developing XML technology. Structoids will be implemented as XML and will have XML Schemas. We are also exploring the use of XML for the definition of Digital Object behavior interfaces and as a means of persistent storage of Fedora content (on the back end of the repository). We hope to take advantage of pattern matching and/or inference of types facilitated by the imposition of XML structure on content. We have also conceived of a Fedora batch loader that uses an XML schema for input. And lastly, we have done some initial experimentation with XSLT as a means to deliver Fedora content. There is some interesting similarity between the notion of "transformation" in XSLT and the "transformation" of digital content to disseminations in Fedora.

We have also created a prototype of web browser public access to Fedora content. This was motivated a) to exploit the nearly ubiquitious HTTP and web browsers and b) to join the web-centric approach that is so common in other Digital Library research. We experimented with SOAP (the Simple Object Access Protocol) for HTTP access, but upon discovery that web browsers are not yet able to speak SOAP, we instead created some java Servlets to provide Web browser access. We restricted ourselves to Fedora public access methods only due to the difficulties of a) associating binary content with text or other binary content in HTTP and b) representing the object-oriented Fedora model in the stateless HTTP environment. In addition to using web browsers for further Fedora client work, we intend to explore how much we can leverage "cookies" and "sessions" as used by java Servlets in our security work.

2 Digital Preservation

A Cornell University Library research team consisting of Anne R. Kenney, Oya Y. Rieger, Peter Botticelli, and Richard Entlich leads investigation of policies and mechanisms for digital preservation. Digital preservation research is also being undertaken by James Cheney, a PhD student in the Computer Science Department, under the direction of Bill Arms, Peter Botticelli, and Carl Lagoze. The preservation team is investigating digital preservation issues in the following research areas.

2.1. Defining an Ideal Preservation Service for a Research

This white paper will offer a high-level conceptual view of an ideal preservation service for a research library operating in a networked environment in which control over important content and services is distributed. Key questions to be addressed in the report are:

- What aspects of digital objects need to be preserved? What is an acceptable level of preservation (minimal and ideal requirements)?
- What are the suites of services that make up a Preservation Service?
- What is the proper balance between computer-enforceable and human-enforceable components?
- How do we define "information loss"?

In addressing these questions, the white paper will focus on preservation issues and resource requirements that are distinctive of research libraries. It will suggest ideal responses, and then assess the status in developing those capabilities (i.e., already in place, here but not implemented in this context, near term, on the horizon, not within our lifetimes). The authors are conducting research and participating in key committees in preparation for this report. Anticipated completion of the white paper is late spring 2001. The Council on Library and Information Resources (CLIR) has expressed interest in publishing this white paper in the summer 2001.

In support of this research, members of the research team participated in a number of international activities:

- Oya Rieger served on the Research Libraries Group/Digital Library Federation Task Force on Policies and Practice for the Long-Term Retention of Digital Materials. The task group gathered and analyzed existing digital preservation policies and practice descriptions to develop an outline for best practices in policy development for libraries.
- Anne Kenney serves on the RLG/OCLC Working Group on Digital Preservation: Defining a Sustainable Digital Archive. This international committee has defined its goals, identified an international membership, and agreed upon key resources to inform its work. It is anticipated the Working Group will be actively engaged in the development of requirements in the coming months.
- Anne Kenney represented Project Prism at the NEDLIB invitational workshop in December on "Setting up Deposit Systems for Electronic Publications (DSEP),"

held in the Netherlands at the Koninklijke Bibliotheek (KB) in The Hague. The NEDLIB project (Networked European Deposit Library) is a collaborative project of European national libraries to build a framework for a networked deposit library, and is one of the key preservation projects that the PRISM staff in monitoring. The project adapted the OAIS reference model to incorporate a preservation component as the basis for the Deposit System, and tested the use of emulation-based strategies for preservation.

- 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.

2.2. Digital Information Longevity Study

There is anecdotal evidence that cultural and educational institutions are losing digital information due to technical and organizational threats, including obsolescence of various technical components, incomplete documentation, and lack of resources dedicated to preservation. However, there are no systematic quantitative studies that document the extent and rate of loss. The goal of the longevity study is to analyze and document loss and evaluate its significance to the usability and value of digital information. A better understanding of leading causes of data loss will help to prioritize preservation strategies and will facilitate risk assessment and the development of preventive measures. As a part of this study, the accomplishments of the group heretofore include:

- Developed metrics to characterize preservation risks from different perspectives (collection manager, user, resource). This exercise highlighted the challenges associated with defining loss and developing formal metrics for its assessment.
- The library team worked with Professor William Arms to test a Web profiler tool that was created specifically to track preservation-significant changes to open access Web sites. This investigation informed the team about usability and effectiveness of automated tools in digital preservation (especially in monitoring and diagnosis).
- After exploring the challenges associated with using automated tools for gathering integrity-check information, the group decided to focus the longevity study on early electronic publications. Using catalogs of electronic resources for research institutions for the past decade, the team is devising a methodology to track the lifecycle of a sampled subset of early digital information resources. The goal is to collect quantitative and qualitative data to better understand how electronic resources have changed and evolved since 1991 (starting with Telnet and Gopher based information resources). In addition to the hands-on assessment of a sample of electronic resources for various integrity parameters (continuity, interface changes, quality control guidelines, added-value services, etc.), the publishers of a selected group of publications will be interviewed to identify in-house organizational processes and policies relevant to digital preservation.

2.3. Preservation Metadata

In this track of research, the goal is to analyze the role of preservation metadata in ensuring information integrity in a distributed digital library context. The findings of the preliminary requirements analysis study (which investigated the existing policies, concerns, and needs in regard to preserving digital content) confirmed the strong relationship between administrative and technical aspects of preservation, stressing the importance of avoiding viewing preservation as a heavily mechanical component. Preservation metadata needs to stimulate action to support a number of pragmatic, managerial, philosophical, and computational functions to fulfill the following requirements:

- Detect preservation-related threats and take action for protection or recovery
- Facilitate preservation decision making and administration through metadata knowledge management
- Promote preventive measures to control and minimize risks

The functionality of a preservation system (as a suite of services) heavily relies on the existence of an extensible preservation metadata model. For example, to be able to monitor and detect factors that would threaten their integrity, digital objects need to be equipped with metadata that will help them to identify events that need to trigger preservation action. Metadata in this context is seen as a facilitator and a catalyst (i.e., similar to rights management metadata that assists in filtering users based on certain criteria), not a passive mass of administrative information recorded in a database. One of the intellectual exercises of the Project Prism team was to examine and compare the existing preservation metadata models. This research track has been instrumental in the development of the longevity study.

As part of this research track, Oya Rieger co-chairs the ANSI/NISO Technical Metadata for Digital Still Images Standards Committee (http://www.niso.org/commitau.html). Technical metadata, which describes various aspects of image characteristics and the capture process, is increasingly being perceived as an essential component of any digitization initiative to ensure the longevity of digital collections. Image metadata work to date within the library and cultural heritage community has focused on defining descriptive elements for discovery and identification. The goal of the NISO Technical Metadata for Digital Still Images Standards initiative is to fill this gap by developing a generalized technical metadata standard applicable to all images regardless of their method of creation. The ultimate goal of the standard is to facilitate the development of applications to validate, process, manage, and migrate images of enduring value. Rieger also participates in the OCLC-RLG Preservation Metadata Working Group. Within the context of the Open Archival Information System reference model, the group aims to develop a set of "essential" preservation metadata elements, drawing on the work of current digital archive projects and metadata experts.

2.4. Organizing Digital Libraries for Preservation and Access: A Study of Digital Library Federation Partners

With recent growth in digital collections, research libraries have begun to face significant management challenges in creating and preserving digital files. As we learn more about the technological risks to digital holdings, libraries are beginning to see the inherent danger in relying heavily on grant-funded projects. Thus, many institutions are planning to create permanent digital library programs, as well as to integrate digital activities more closely into established library functions. In general, it is becoming clear that to ensure the long-term integrity of digital materials, libraries will have to reallocate permanent resources and develop new organizational structures to meet future digital preservation and access needs.

To address these challenges, Project Prism will work with staff from the Digital Library Federation (DLF) and the Council on Library and Information Resources to produce a series of in-depth case studies involving the two-dozen research libraries that are currently partners in the DLF. Data gathering will begin with the results of an ongoing DLF survey of its partner libraries, which will identify and begin to categorize digital library policies and institutional responsibilities. In turn, the Prism study will focus on the basic library functions or services that directly affect the long-term integrity of digital materials. The impact of existing policies and plans will be evaluated by mapping the work processes, or routines, whereby digital materials are created and managed.

The project will also define organizational models based on data from the case studies, in an effort to compare the effectiveness of different strategies in managing known risks to digital collections. It is also hoped that the case studies will show how libraries can develop more flexible organizations that will be better able to adapt to the unforeseeable consequences of future technological innovation. Most importantly, this project assumes that digital library organizations will have to evolve continually as libraries seek to implement new technologies and expand the range of their digital holdings. In the long run, the job of organizing a digital library must be viewed as an ongoing process, with policymakers and managers paying close attention to the changing resource requirements for effective digital preservation and access. Hence, this study is intended to serve as a first step in an ongoing program for monitoring changes in the organizational and technical environment of digital libraries, in an effort to support the work of library managers as they seek to ensure the future viability of digital resources.

2.5. Development of Deposit Guidelines for Digital Image Collections

The Prism team led a library-wide effort to develop recommendations for establishing a central depository for preserving Cornell’s digital image collections. Their report calls for the establishment of a central depository within the library’s Digital Library and Information Technology (D-LIT) infrastructure for ensuring continuing access to digital image collections over time. Such a central responsibility is seen as essential to ensuring a cost-effective preservation capability. The report outlines the requirements for deposit (including content and usability, legal considerations, conversion, formats, storage, and metadata) and identifies the role and major responsibilities of a central depository. The report concludes with recommended next steps, including a three-month feasibility study to determine the library’s readiness to implement such a central depository responsibility.

This report will be submitted to the Library Management Team in February 2001 and made available on the Prism Web site.

2.6. Mathematical Models for Preservation

James Cheney, Carl Lagoze and Bill Arms in Computer Science and Peter Botticelli of the Cornell University Library have been collaborating on describing long-term preservation using mathematical models. By taking this approach, we hope to better understand how to analyze and solve preservation problems from first principles, rather than solely through an evolving body of practice. Such a theoretical understanding might both establish the limits of preservation and guide future preservation practices.

We seek to elevate the understanding of what the "information" is from "just bits" to high-level abstractions that are more representative of what people really care about. These models draw inspiration from information theory and programming language semantics. We distinguish between physical data, like bits stored on a disk, and logical or abstract information content, such as numbers or letters, images, or even programs. These distinct domains are linked by interpretations that describe how information content is represented physically. Physical objects are subject to change from both the physical environment and the actions of agents over time. Based on this framework it is possible to give a formal definition of preservation and to formulate languages of preservation plans and policies, which describe agent behavior and goals, respectively. Currently we are developing this "theory of preservation" further and attempting to validate it by modeling case studies.

A paper on this work is anticipated for submission to the ECDL 2001 conference.

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. Work in this area is led by Geri Gay, with assistance from Jenna Burrell, Michael Grace-Martin, Helene Hembrooke, and Michael Stefanone.

3.1. Search Behavior

Research investigating search strategies as a function of topic expertise is currently underway. In our current study we will be addressing how participants describe their search strategies, how these descriptions vary between different expertise conditions and by gender. In addition self report descriptions will be compared to log data of search terms to determine the relative amount of overlap between what participants said they tried and what they actually did. Finally, path analyses and Multidimensional scaling techniques will be applied to both self descriptions and log data in order to visually represent search behavior and how this differs with increases in expertise and information specificity.

3.2. Tracking Behavior Online

Besides the search behavior studies, we are examining log files from approximately 180 students over the course of two semesters. Students' Web browsing on these laptops (including: URLs, dates, and times) was recorded 24 hours/day, 7 days/week in a log file by a proxy server during most of a semester (about 30 weeks). We categorized the content of the top 2000 URL hosts (in terms of hits) appearing in the over 1.7 million records of Web browsing data we collected in our proxy server log over the course of the entire year. By doing so, we were able to assign a category to approximately 87% of the URLs we captured regarding test students' browsing. For each student, browsing behavior was quantified and then correlated with academic performance. The emergence of statistically significant results suggests that quantitative characteristics of browsing behavior-even prior to examining browsing *content*-can be useful predictors of meaningful behavioral outcomes. Variables such as *Number* of browsing sessions and *Length* of browsing sessions were found to correlate with students' final grades; the valence and magnitude of these correlations were found to interact with Course (i.e., whether student was enrolled in the *Communication* or *Computer Science* course), Browsing Context (i.e., setting in which browsing took place: *during class*, on the *wireless network between classes*, or *at home*) and Gender. Finally we found that the relative prevalence of social computing *increased* and became more *exclusive* for students in the communication course, especially on the *wireless* network. Social computing and use of the wireless network were *less* prominent and influential for students in the computer science course.

3.3. Context Aware Computing in Museum and Library Settings

Freeing users from the desktop is now a practical reality in many environments. The implications for mobility are both far-reaching and under-realized in many of the current scenarios we have seen. Our work has focused on the integration of user input into the iterative design process used to develop a contextually aware application for use in library and museum settings. Our research has looked at two new models of computer usage that pertain specifically to mobile, networked devices. The first model is context-aware computing. This is a networked environment where information is made available to a user dependent upon specific environmental factors. These factors can include location, time, user and device type. Our second model encompasses social navigation. Social navigation integrates the experiences of previous users, making that knowledge available to the currently connected user. In unifying these two models a connected device can make visible what is invisible in the users environment by drawing upon the collective knowledge of previous users. This knowledge is made available to indicate what is relevant and interesting about the physical space or the task at hand.

4 Interoperability Architectures

During the past year we made substantial progress in developing and disseminating the work of the Open Archives Initiative (OAI) (http://www.openarchives.org). The goal of the OAI is to provide a collaborative framework for development and maintenance of low-barrier digital library interoperability solutions. Carl Lagoze and Herbert Van de Sompel lead OAI work at Cornell. Additional organizational support for OAI comes from the Digital Library Federation and the Coalition for Networked Information.

As described in last year's annual report, the OAI began in the EPrints community. In a meeting held in late 1999, members of the EPrint community, and other interested parties, met to develop practical solutions for federating EPrint archives. The result was the so-called *Santa Fe Convention*, a protocol for harvesting metadata in multiple formats. The basis for the protocol was the Dienst protocol, developed out of DARPA-funded work at Cornell. By the end of the previous reporting year, a number of the participants in the Santa Fe meeting had committed to experimentation with the components of the Santa Fe Convention.

In 2nd quarter 2000, after dissemination of the Santa Fe Convention, it became apparent that the idea of a low-barrier technical framework based on metadata harvesting had appeal beyond the EPrint community. This evidence came out of two OAI workshops organized by Carl Lagoze (in partnership with Ed Fox at Virginia Tech) at the ACM DL workshop in San Antonio and the ECDL workshop in Lisbon. Further evidence of broad interest came from a set of meetings sponsored by the Digital Library Federation and the Andrew W. Mellon Foundation, where members of the research library community committed to this approach.

The outcome of these events were a number of important changes in the OAI:

- A shift in mission from an EPrint focus to one more applications neutral, acknowledging the wider application of the technology.
- The formation of an internationally based steering committee, recognizing that dissemination of standards relies on a stable organization.
- Organizational support from the DLF and CNI to Cornell for maintenance of the developing standard.
- The formation of a technical committee which met over a period of two days in September 2000 at Cornell to modify and refine the technical infrastructure in light of the experiments heretofore and the broadened application framework.

The culmination of these changes was intense development work during the fourth quarter of 2000 on the complete specification and testing of the Open Archives Initiative Protocol for Metadata Harvesting (http://www.openarchives.org/OAI/openarchivesprotocol.htm). This development work was led by Cornell but included a variety of participants (in both technical development and alpha testing) including Los Alamos National Laboratories, Library of Congress, OCLC, CIMI, and UKOLN. The result is a technical infrastructure that has been vetted and shown useful by a wide variety of communities.

Public release of the protocol occurred in two "open meetings" sponsored by the Digital Library Federation. The first for US participants occurred on January 23, 2001 in Washington DC. The second, for European participants occurred on February 26, 2001 in Berlin.

The public release of the protocol begins a 12-18 month period of controlled experimentation. Our intention over this period is to keep the protocol stable in both scope and substance and determine if the notion of metadata harvesting, and the extensibility mechanisms built into the protocol, are sufficient for building a useful service infrastructure. Our work at Cornell during this period will consist of managing this experiment including:

- maintaining a registry of data repositories that support the OAI protocol.
- periodically verifying protocol compliance among registrants to ensure integrity of the protocol.
- working to create an "open archives community" of data providers and service providers (clients of the protocol) to encourage broadest testing of the concepts.

Follow up meetings to determine results of the interoperability experiment will be scheduled in first or second quarter of 2002.

5 Policy Enforcement

A key part of our Prism research continues to be in the area of *policy*, including the definition, formal declaration, and enforcement of policy. A vital part of digital library integrity is the ability to create, support, and enforce policies that define access management and preservation. Sandy Payette leads policy enforcement work on Project Prism, with assistance from Fred Schneider and Vicky Weissman. This work is jointly funded by Cornell's Information Assurance Institute.

The issue of *policy specification* (definition and formal declaration) spans the terrain from the point where humans intellectually define policies, through the formalization of those definitions into policy declarations, to the refinement of those declarations into specifications that can be understood by automated enforcement mechanisms. We are working on several Prism sub-projects in this area to examine this full spectrum/lifecycle By collaborating across these projects we are investigating the optimal process for maintaining integrity in the translation between these different manifestations of a policy. If integrity is not maintained, a policy may not be enforceable, or it may be enforced in a manner that violates the original intention of the humans who conceived of the policy.

The issues of policy specification are intimately tied to the challenges of *policy enforcement.* In the digital library domain, a basic requirement for both policy specification and enforcement is the ability to accommodate a wide range of digital objects and usage scenarios. We need specification languages that are highly expressive, and enforcement mechanisms that are flexible and extensible. This stands in contrast to traditional access control models, which are limited to a relatively fixed set of operating system abstractions (e.g., files) and computing actions (e.g., read, write). In the area of policy enforcement, we are investigating the use of new, language-based security techniques to provide a richer and more extensible form of policy enforcement for complex digital objects in distributed digital libraries. This work is grounded in Schneider's theory of security automata, which forms the basis for the policy enforcement technique known as In-Line Reference Monitoring (IRM). [Schneider, July 1999; Erlingsson and Schneider, July 1999].

Within this framework, we are currently working on a number of specific projects, detailed below.

5.1. Policy-Carrying Policy-Enforcing (PCPE) Digital Objects

In this project we are experimenting with an object-centric model of policy enforcement that involves locating policies within the digital objects to which they pertain. Also, by using In-line Reference Monitors (IRMs), our digital objects are able to perform their own policy enforcement. Last year we developed an initial prototype of the PCPE concept using our Fedora software and Cornell's PoET software for IRM policy enforcement. This year's accomplishments in this project include:

- *Refinement:* we significantly evolved the technical aspects of our prototype
- *Testing*: we challenged our prototype with more complex digital objects, and more complex policies.
- *Implementation*: this work has been permanently integrated in our Fedora reference implementation software.
- *Generalization of design*: we generalized the design of our prototype so that we can now specify a means of interfacing an extensible application (such as a Fedora repository) with a mechanism that enables dynamic IRM policy enforcement. We have taken this design forward into the next phase of our work, described below in *Policy Intention Architecture*.

Additionally, this project has successfully demonstrated the following benefits of the PCPE approach:

- Policies can be completely tailored to meet the needs of specific items, without over-burdening a system-wide mechanism with idiosyncratic policy rules
- Objects can be extensible in their functionality, and policies can be modified to reflect new or changed behaviors
- Objects that contain their own policies are comprehensive units that can be managed over time by their authors or stewards, instead of by system managers.
- Objects can be moved among trusted repositories or to portable devices without losing their customized policies.

5.2. Policy Intention Architecture

When a human expresses a policy and expects that policy to result in a certain outcome, we say that that human has a *policy intention*. A policy intention is violated when the policy is not enforced, or when the policy is enforced in an unanticipated or undesired manner. A policy intention recognizes the hand-in-glove relationship between policy specification and policy enforcement. The PIA is being designed to ensure integrity between specified policy intentions and the resultant outcome of enforcement - with a particular focus on IRM enforcement.

This year's accomplishments in this area are outlined in the following sections.

5.2.1 Analysis of the Policy Specification Process

We are investigating the process of evolving a policy from a human conception to an IRM-enforceable policy. This process can be generally characterized as: (human articulation → formal declaration → intermediary specification → IRM specification). This effort is done in collaboration with other related Prism sub-projects in both the research library and computer science context.

5.2.2 Proof-of-Concept for Dynamic Policy Intentions

We developed an early prototype to test the concept of declaring policies that speak about specific *properties* of complex digital objects. By properties, we mean any significant attribute of a digital object -- like the fact that the object is of a certain type (a book or a lecture), or that it is authored by a particular person, or even that contains certain information (the word "security" appears somewhere in the object). If a digital object is able to expose a view of its inherent properties in a normalized manner, then we can talk about these properties in our policies. If our policies can refer to objects by their properties (as opposed to just their unique identifiers), then our policies can be very dynamic -- they can refer to a *set of objects* that meet certain properties. The goal was to demonstrate this concept. We developed a proof-of-concept prototype using XML/XSLT. Essentially, we wrote policies that used XPATH expressions to specify patterns that could exist in one or more digital objects. A policy statement was a set of patterns (a constraint on the domain of all digital objects) and a set of associated restrictions. Given that Fedora Digital Objects were able to expose a view of their properties in XML, our prototype was able to identify an object, or a subset of objects, to which particular restrictions applied. These restrictions were fine-grained in that they restricted access to particular methods that could be run on the Digital Object (like the ability to view slides on lecture objects). Our successful demonstration of policy intentions based on object properties has helped frame our design of the *Policy Intention Registry (PIREG)* component of our emerging architecture for policy enforcement (see below).

5.2.3 Development of an Extensible Schema-Based Policy Language (ESBPL)

We are developing a policy specification language that will specify an intermediary format - one that resides between a highly abstract, general-purpose policy language (see 2c) and a specific format optimized to support bytecode inlining (IRM). ESBPL provides a general means for encoding declarations about objects, subjects, runtime environments, and restrictions. Recognizing that different communities and application environments define these things differently, ESBPL uses a nested schema-based approach to ensure flexibility and extensibility. We are implementing ESBPL using XML and XML schema language.

5.2.4 Design of the Policy Intention Architecture (PIA)

We are designing an architecture that facilitates IRM policy enforcement for repositories of complex, extensible digital objects. We plan for this work to generalize to other types of extensible applications. Our intent is to support the policy specification process (with a module that interoperates with logic-based policy specification tools (as described in 2c). The PIA supports IRM policy enforcement by interfacing with Cornell's PoET software. A key module in the PIA is the Policy and Code Manager (PCM), which orchestrates the associating of user-articulated polices with bytecode modules. In summary, the components of the PIA are:

- *Policy and Code Manager (PCM)* - this component is the runtime interface of the PIA. If an application wants to be IRM-enabled it will need to plug into the PCM. The job of the PCM is to figure out what pieces of code need to be in-lined with which policies - based on runtime conditions (as requests are being made within the application domain). The PCM will not let any piece of code run unless it is properly protected (i.e., the code must be in-lined with the "right" policies before that code is allowed to run).
- *Policy Intention Registry (PIREG)* - the database of all policies for a repository, encoded in the ESBPL described above. We are planning for the PIREG will interface with the logic-based specification module described below.
- *PoET* - Cornell's toolkit for IRM policy enforcement.

5.2.5 Testbed definition and design

Our first investigative model is centered around policy problems encountered in online distance education applications. This application area provides a cross between a process-oriented policies and typical access control policies for digital collections. We have specified test scenarios that deal with the processes of course delivery and consumption. We will focus on fine-grained policies required to manage access to multiple digital collections as users progress through the processes of course delivery and course consumption. We will also focus on the architectural issues (interoperability and extensibility) to ensure that the PIA is plug-and-play and can support different kinds of applications.

5.3. A Logic for Policy Specification

Our goal to design a logical framework to support policy expression and analysis in dynamic distributed systems. Over the past decade, several logics have been proposed; each allowing almost any conceivable policy to be expressed. Unfortunately, this flexibility is more than an application needs and results in a logic for which most questions, such as consistency, are undecidable. In the digital library context we aim to discover a class of policies that are relevant to a real application. In fact, the policy set will be interesting to a number of systems that share the concerns of the digital library community. We can then refine a previously proposed logic; exchanging the ability to express policies outside of the class with the ability to reason about policies within the class.

The project's current status is as follows. Sample policies have been collected from the digital library community. Based on these samples, an interface has been designed and implemented to allow digital content managers to specify policies. By using this interface 'in the field', we will discover the useful policies that cannot be captured in the current restricted setting and modify the structure accordingly. Policies entered through the interface are translated into a subset of the logic proposed by Joseph Y. Halpern, Ron van der Meyden and Fred B. Schneider in "Less is More: Logical basics for Trust Management." Analysis of this subset is underway.

6 Web Preservation

The Web Preservation Project at the Library of Congress began in March 2000. Preserving open access materials from the web falls within the Library of Congress's mission to collect and preserve the cultural and intellectual artifacts of today for the benefit of future generations. The report by the National Research Council, *A Digital Strategy for the Library of Congress* (Carl Lagoze served on the NRC committee to produce this report), urged the Library to move ahead rapidly. William Arms is advising the library on its strategy and, during 2000, coordinated a prototype.

6.1. The Prototype

The main activities in developing the prototype were as follows:

- A small number of web sites were nominated by selection officers at the Library of Congress. Of these, 35 were examined and three sites were chosen for close study.

- Snapshots of these sites were downloaded using the HTTrack mirroring program. The snapshots were inspected for errors, anomalies, etc.

- Catalog records were created using OCLC's CORC software and loaded into Library of Congress's ILS system.

- A trial web site was developed to evaluate user access. This allowed users to view each version of each website as it was downloaded. The site also allowed users to search the ILS, or lists of URLs, titles and subject headings, with links to the collection of web sites.

6.2. Strategic Directions

In December 2000, a report was issued to the library describing the pilot phase of the Web Preservation Project and making recommendations about collecting and preserving such materials for the long term. This report covers the following topics:

- Strategies for collecting open access web materials. The two major approaches are bulk collection, which is entirely automatic, and selective collection, carried out by skilled librarians.

- Use of the collections for scholarship and research. The choice lies between analysis of raw snapshot files by computer program and human analysis using a rendered version of the preserved web site. The second requires that the snapshot file be edited to create an access version that can be rendered successfully.

- Information discovery. The report discusses the choices for information discovery. The options include catalog records, either at item level or collection level for groups of sites. The records themselves can be full MARC records or use some shortened form, such as Dublin Core. Other means of access that might be provided at much lower cost include searchable lists of URLs, <title> fields extracted from HTML pages, or free text indexes of home pages.

- Legal issues. A series of meetings with members of the Copyright Office have identified the necessary changes in copyright regulations and laws that will permit the Library of Congress, with partners, to carryout its duty to preserve materials downloaded from the web.

- Long-term preservation. The report examines the options for long-term preservation. It recommends a combination of keeping refreshed versions of the original snapshots with automatic migration of files to modern formats. It is not possible to guarantee that the experience of using web sites will be preserved.

- Development of a production system. The report includes rough estimates of the volumes and costs of a productions system.

7 Publications

Burrell, J., Treadwell, P., and G.K.Gay. Designing for Context: Usability in a Ubiquitous Environment. In *Proceedings of the 2000 Conference on Universal Usability*, 2000.

Dushay, N. and C. Lagoze, Modeling Decisions for Digital Content, Computer Science Technical Report TR2000-1807. 2000, Cornell University Computer Science.

Gay, G.K., M. Stefanone, M. Grace-Martin, and H. Hembrooke. (in press). The Effects of Wireless Computing in Collaborative Learning Environments. *International Journal of Human-Computer Interaction*.

Grace-Martin, M. & G. K. Gay. (in press). Web Browsing, Mobile Computing and Academic Performance. Special Issue on Cirriculum, Instruction, Learning and the Internet. *IEEE and International Forum of Educational Technology & Society*.

Jones, M. L. W., R. H. Rieger, P. Treadwell, and G. K. Gay. Live from the Stacks: User Feedback on Mobile Computers and Wireless Tools for Library Patrons. *ACM Digital Libraries 2000,* 2000. Berkeley, CA.

Kenney, A.R. and O. Rieger, *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View, CA: Research Libraries Group, Inc., 2000.

Kenney, A.R. and O. Rieger. Preserving Digital Assets: Cornell's Digital Image Collections Project. *First Monday* 5:6 (2000).

Kenney, A.R. Collaborating Across Lines: Librarians, Archivists, and Computer Science Researchers in Cornell's Project Prism. *Society of American Archivists* annual meeting, August 31, 2000, Denver, CO.

Kenney, A.R. Preserving Cornell's Digital Collections. *Web-Wise: A Conference on Libraries and Museums in the Digital World*, March 15-17, 2000, Washington DC.

Kenney, A.R., Mainstreaming Digitzation into the Mission of Cultural Repositories. *Collections, Content and the Web*. Washington DC: Council on Library and Information Resources, 2000.

Kenney, A.R., Projects to Programs: Mainstreaming Digital Imaging Initiatives, in *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View, CA: Research Libraries Group, Inc., 2000.

Lagoze, C. and H. Van de Sompel. The Open Archives Initiative: Building a Low-barrier Interoperability Framework. Submitted to *Joint Conference on Digital Libraries*, 2001, Roanoke, VA.

Lagoze, C. The Cornell Digital Library Research Group: Architectures and Policies for Distributed Digital Libraries. *DLW17,* 2000, Tsukuba, Japan.

Lawrence, G., W. Kehoe, O. Rieger, W. Walters and A.R. Kenney. Risk Management of Digital Information: A File Format Investigation. Washington, DC: Council on Library and Information Resources, 2000.

Payette, S. and C. Lagoze, Metadata: Principles, Practices, and Challenges, *in Moving Theory into Practice: Digital Imaging for Libraries and Archives,* Research Libraries Group, Spring 2000

Payette, S. and C. Lagoze, Value-Added Surrogates for Distributed Content: Establishing a Virtual Control Zone. *D-Lib Magazine*, 2000. **6**(6).

Payette, S. and C. Lagoze. Policy-Enforcing, Policy-Carrying Digital Objects. *Fourth European Conference on Research and Advanced Technology for Digital Libraries*. 2000. Lisbon, Portugal.

Payette, S., et al., Interoperability for Digital Objects and Repositories: The Cornell/CNRI Experiments. *D-Lib Magazine*, 2000. **6**(5).

Rieger, O. Project Prism: Preservation Metadata Research. *Information Infrastructures for Digital Preservation Conference*, December 2000, York, UK.

Rieger, O. Projects to Programs: Developing a Digital Preservation Policy, in *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View, CA: Research Libraries Group, Inc., 2000.

Van de Sompel, H. and C. Lagoze, The Santa Fe Convention of the Open Archives Initiative, *D-Lib Magazine*, 2000. **7**(2).

Van de Sompel, H and C. Lagoze (ed.), The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 1.0, 2001.