Fedora and the Preservation of University Records
Interim Narrative Report

NHPRC Grant #2004-083
January 31, 2005

Submitted by:

Eliot Wilczek
Digital Collections and Archives
Tufts University

Kevin Glick
Manuscripts & Archives
Yale University

The Digital Collections and Archives (DCA), Tufts University, in conjunction with Manuscripts and Archives (MSSA) of Yale University Library, is working on a National Historical Publications and Records Commission (NHPRC) electronic records research grant project (2004-083) to synthesize electronic records preservation research with digital library repository research in an effort to develop systems capable of preserving university electronic records at both institutions. This project is testing the potential of Fedora (the Flexible Extensible Digital Object and Repository Architecture) to serve as the architecture for such an electronic records preservation system. The project's original grant proposal narrative and the revised plan of work can be found on the project's website at: <http://dca.tufts.edu/features/nhprc/reports/index.html>.

Since the project's beginning in July 2004, much work has been undertaken in just a few months time. Administratively, we established the projects formally at both research institutions. A website <http://dca.tufts.edu/features/nhprc/index.html> was created and is maintained by Tufts to serve as the public face of the project. Notice of the project was distributed through appropriate professional email lists as well as through university and library press releases. The distance separating both sets of researchers necessitated utilizing an online document management system. With funds shared by Tufts and Yale, the project team has implemented Groove Virtual Office to help us work together securely over the Internet simultaneously. The software allows us to share information, manage project tasks, conduct meetings and get work done. Two face to face meetings of the entire group were conducted in this period, one at New Haven in August, and another at Tufts in November. Activities to disseminated the research findings are planned for the near future with presentations at ECURE in March, ACRL-NEC in April, and at the Fedora Users' Conference in May. In addition, the project was briefly mentioned in a presentation on institutional repositories in January.

A change to the make-up of the project staff has lead to a different outlook and some very productive work. We have been fortunate to bring in some very skilled computer professionals. We hired Robert Dockins to be our Fedora Project Analyst and he began working in late August. Mr. Dockins is a highly skilled and motivated programmer and has previous Fedora experience. Since we submitted our original grant proposal in May 2003 there have been several administrative changes at the Yale University Libraries that made it necessary to reconfigure some of the grant project staff at Yale. Stephen Yearl, Systems and Digital Resources Archivist

in Manuscripts & Archives was not able to devote 25% of his time to the project. Instead, we have split his time, all in-kind contributions, with Raman Prasad, Computer and Information Systems Support Specialist in Manuscripts & Archives. Mr. Yearl and Mr. Prasad have therefore each been contributing 12.5% percent of their time to the project. In addition, David Gewirtz, Project Manager of the Yale Library Project, Academic Media & Technology, Yale ITS, was not be able to devote 50% of his time to the project. In his place, Roy Lechich, Senior Programmer Analyst, Library Systems Office is contributing 20% of his time; Xinjian Guo, Systems Programmer, Library Project, Academic Media & Technology, Yale ITS is contributing 25% of his time; and Mr. Gewirtz to devoting 5% of his time. This is all NHPRC sponsored funds. Mr. Lechich and Mr. Guo are skilled and expert programmers who have been working alongside Mr. Gewirtz on Yale Library systems projects. They have been tremendous assets to the project and alongside our more technical additions of Raman Prasad and Robert Dockins at Tufts, we have been able to refocus our efforts more toward building specific tools to support the accessioning of electronic records into a preservation repository. Such tools are necessary to make Fedora more suitable for preservation of university electronic records and more compliant with OAIS model specifications.

This change of staff has also resulted in a slight change in the order to the project's activities. Much more time has been spent on what was originally planned to be the third phase of the project, where we would articulate the process of ingesting electronic records into a preservation repository. Our original plan was to outline and describe the steps in the process and to eventually attempt some test accessions to evaluate our descriptions and to evaluate Fedora. We have moved beyond this initial goal to more technically break down each part of the process into a detailed list of steps and corresponding tools needed to complete those steps. We have focused much of our attention specifically on the transformation of an archival submission into an archival information package that can be stored in Fedora. According to the *OAIS Reference Model*, this activity includes, "Transforming one or more SIPs into one or more AIPs that conform to the archive's data formatting and documentation standards. This may involve file format conversions, data representation conversions or reorganization of the content information in the SIPs. The Generate AIP function may issue report requests to Data Management to obtain reports of information needed by the Generate AIP function to produce the Descriptive Information that completes the AIP. This function sends SIPs or AIPs for audit to the Audit

Submission function in Administration, and receives back an audit report." A preliminary look at this process identifies that this activity is initiated by the arrival of a SIP in the "drop box"—the SIP must include the content file(s), the submission agreement (although this may appear before the SIP), producer-supplied metadata (such as descriptive, authenticity, and provenance), and the manifest (including bit-level checks). Generally, the steps in this activity include:

- Validate the SIP and Verify Success of the Transfer [virus check packed files, bit-level check of packed files, unpack, virus check unpacked files, bit-level check of unpacked files (if necessary), programmatic check for 4 parts (valid SA, files in SIP match those described in manifest, etc)].
- Verify Authorization to Transfer [check SA against a list of producers authorized to submit SIPs].
- Register Transfer (serial check-in). This would depend on the naming service.
- Register if Authenticity can be presumed.
- Validate Records Components (using validation tools that depend on file types, e.g. JHOVE, OO modules, etc.).
- Validate Producer Created Metadata.
- Register if Transformation is Necessary for Ingest (Both records components and metadata). This judgment is based on the feasibility of preservation assessment included in the SA, Preservation Strategies and Standards external to ingest (from Preservation Planning), Format and Documentation Standards and Data Management and Archival Storage Procedures (from Administration).
- Formally Accept Records (acknowledge submission to producer, document acknowledgement).
- Transform SIP to AIP. The AIP includes:
    - Digital content (in original format) of the record components
    - Digital content (in transformed format) if transformation undertaken
    - Representation information to understand and view the format
    - Confirmation of authenticity and data integrity
    - Producer-supplied descriptive metadata (to facilitate access)
    - Administrative metadata (provenance info, accession record, preservation action record, appraisal decisions, etc.)
    - Structural metadata to document relationship among files

The new focus on the technical aspects of each part of ingest process—particularly transformation of submission information packets to archival information packets—and the evaluation and development of the tools required by these aspects of ingest have delayed the project's completion of Steps One and Two. In addition, Step One—articulating functional requirements for active recordkeeping systems that will transfer records to an archival repository—and Step Two—articulating functional requirements for archival repository preservation systems have proved to be more complicated tasks to complete than initially anticipated.

For Step One, the ten reports of functional requirements we are analyzing often have inconsistent definitions of "recordkeeping system" which has prompted us to spend a significant amount of time carefully defining this term. We feel that this careful defining of recordkeeping system will facilitate the creation of a detailed and usable compilation of requirements for recordkeeping systems. Because the existing requirements for preservation systems is not at as rich the requirements for recordkeeping systems, Step Two has turned out to be much more dependent on Step One than we had originally anticipated. Despite the delay, we still anticipate completing both of these tasks and believe that they will only be improved with the additional time spent on their development.

At this point, the project is proceeding successfully, well on its way toward achieving its objectives in a timely manner.