INTERIM REPORT OF THE INSTITUTIONAL REPOSITORY POLICY COMMITTEE

1. Executive Summary

Digital repositories appear to be one promising way to archive, protect, and simultaneously make widely available via the Internet, intellectual property created by University faculties. In September 2003, Associate Provost for Research Dennis G. Hall convened a faculty committee to examine the concept of an electronic repository to house materials created by Vanderbilt faculty in the course of research and teaching. However, this challenging undertaking requires that we rethink almost every aspect of the ways in which this intellectual property is disseminated. This report summarizes the deliberations of the Institutional Repository Policy Committee (membership appended) during the academic year 2003-2004. It reports on the current status of various repository software projects underway around the world; assesses potential uses of an institutional repository at Vanderbilt; proposes a draft policy on the organization and governance of such a repository; outlines the complex issues surrounding the practical implementation of the repository; discusses hardware requirements and cost projections for the physical repository; and proposes a timeline for implementation that will permit us to develop both capacity and capability for moving forward, while not committing resources prematurely. Our aim is to make the repository a campus-wide product and a world-class resource. The Committee looks forward to the response from the Vanderbilt administration to these deliberations, and to the dialogue that will begin with that response.

2. Background and Perspective

Documents are stored within an institutional repository to (1) permanently preserve scholarly (and possibly some administrative) materials in electronic form; (2) present the face of the University to the world; and (3) make research accessible across different University communities. The driving forces for digital archives include, but are by no means limited to, the skyrocketing costs of commercial scientific publishing, the economics of scholarly presses, and the proliferation of open-access journals and other non-traditional forms of scholarly output, especially at the rapidly evolving boundaries between the canonical disciplines.[1] Powerful software engines are evolving in response to the interests of potential users in developing institutional repositories capable of housing the scholarly output of faculty across all subject areas and in many different data formats..[2] However, ,many complex legal, infrastructural, administrative and financial issues must be faced if the institutional repository concept is to be extended to practice.[3]

DSpace is open-source archival software, based on the Open Archive Movement, which makes it possible both to archive and to access materials of many different data types and formats. At present, DSpace provides a simple interface that (1) sets up communities with leaders who are authorized to determine what materials are included in the community's archive, and (2) collects a common set of basic metadata for the objects in the archive. The primary advantages of setting up a DSpace community as opposed to placing material on an existing server are first, that DSpace materials will be maintained permanently by the Library and migrated to formats that remain usable in the future, and second, that the common metadata system facilitates searching and usability of the items stored worldwide in DSpace archives. Vanderbilt Library has already received the Provost's approval to move forward with a pilot program.

At Vanderbilt, DSpace could work as a repository for course-related documents, easily moving items between the On-line Access to Knowledge [OAK] course management software and DSpace. Also, some VUSpace documents might potentially be appropriate for DSpace, making them searchable using standard metadata descriptions. While DSpace is possibly the most widespread of the Open Archive software packages at present, and we can imagine many potential uses at Vanderbilt, it is still early in its development and we need to adopt a strategy of moving slowly and monitoring software development efforts.

[1] Dennis G. Hall, "Some Thoughts on Scholarly Publishing in the 21st Century," *Optics and Photonics News* 14 (10), 30-33 (2003).
[2] Vivien Marx, "In DSpace, Ideas are Forever," New York Times, August 3, 2003, Education Section.
[3] Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL* 226, 8 (February 2003).

3. Status of Institutional Repository Software Projects

In recent years, development of institutional repositories has taken off rapidly in the U.S. and in Europe. Between fifty and eighty members of the Association of Research Libraries [ARL] are working on institutional repositories, including the University of California, which has established communities for almost every research center, with over 3000 articles and hundreds of books already on line. The UC e-scholarship repository may be accessed at http://repositories.cdlib.org/escholarship/. In England, scholars are using E-Prints, which works exclusively with print archives; at the University of Virginia, there is an open-archive project called Fedora; in Australia, Greenstone is being explored; and at Vanderbilt we are looking into DSpace, a software product developed at MIT with the support of Hewlett-Packard.

The DSpace software has been downloaded more than 8,000 times since its release in summer of 2003. Twenty production sites now operate worldwide, and another 125 institutions, almost equally divided between the United States and the rest of the world, are evaluating the software. The leading research universities using DSpace are: Cambridge, Columbia, Cornell, MIT, Ohio State, Rochester, Toronto and the University of Washington. These universities have a grant from the Mellon Foundation to advance the DSpace software, although many other universities are joining the effort, including the Online Computer Library Center [OCLC], the leading library cooperative of over 20,000 libraries worldwide. At the first open meeting of the DSpace Federation, there were125 attendees from 50 organizations and 9 countries.

4. Proposed Vanderbilt Policy Governing the Institutional Repository.

4.1. Definition of the DSpace Repository

The Vanderbilt DSpace Repository is a partnership between Vanderbilt communities, Vanderbilt Library and Vanderbilt administration. DSpace content will consist of collections produced by Vanderbilt communities, which are managed, preserved and distributed by Vanderbilt Libraries through DSpace. As in all partnerships, it is important that all DSpace stakeholders understand and agree to the policies required to build a DSpace repository.

A fundamental policy question is whether the Library should maintain its traditional role as agent for the University in charge of making selections of materials to acquire, catalogue and archive, or if the faculty should begin to tell the Library directly though the medium of DSpace archives what materials should be preserved. This new technology blurs the line between collector of existing materials and publisher of new materials. The Library neither can nor should preserve every document produced on campus; however, it is not desirable to push the Library into a broad new role as editor or censor. It should also be recognized that DSpace is a complement, not a substitute for VUSpace or Web pages. DSpace is a tool aimed at preserving materials of some general interest that might otherwise disappear. Many things placed on local hard drives or servers, on the other hand, are only intended for temporary or purely personal use. With these considerations in mind, the following policies are proposed.

4.1.1. The Library should take the initiative to create communities, involving appropriate leaders or campus units. Unless the Library takes a proactive role in contacting departments and other groups and fostering the transition to this new technology, many important materials will be lost. Most departments have found at least short-term solutions that meet their immediate needs. They are unlikely to take the initiative to change a working system solely to satisfy the archival desires of the University.

4.1.2. There will also be many types of archive-worthy materials generated by groups on campus of which the Library staff will be unaware. For example, a center or department might put out a newsletter, or the mathematicians might host a conference and wish to preserve the text of the talks that were given, or students might produce a series of short films for a class project that should be archived. The Library should certainly search for such products of the scholarl enterprise and encourage the leaders of these groups to set up DSpace archives, but it should also encourage the campus community generally to form appropriate user communities that will facilitate the overarching objectives of the repository outlined in Section 2.

4.1.3. DSpace archives should try to preserve materials that are likely to be of at least some interest to users in the future. It would be difficult and not very useful to go beyond this principle and attempt to specify in detail which documents should and should not be collected.

4.2. Defining User Communities and Eligible Materials

4.2.1. A DSpace community is a self-constituting group at Vanderbilt that produces research, has a defined scope and long-term stability, and takes responsibility for setting community policies. Each community assigns a coordinator to work with Vanderbilt Libraries staff.

4.2.2. DSpace need not mirror the existing University organizational structure; indeed, it would be surprising if groups with common scholarly interests followed these lines very closely. User communities should both be created exogenously and allowed to arise endogenously.

4.2.3. A Vanderbilt DSpace user community should be any group that: (1) Includes at least some faculty members/administrators associated with the University; (2) Produces secondary materials likely to be of some general interest to users in the future; (3) Produces materials that serve some University mission (research, teaching, administration, etc).

4.2.4. Any faculty member (including non-tenure track and visiting faculty), and any representative of any administrative or other academic organizational unit on campus should be allowed to propose a DSpace community.

4.2.5. Groups wishing to establish a DSpace community that does not conform to this definition will be considered on a case-by-case basis. Individual faculty members may submit items through an established community in the DSpace project.

Note that this definition explicitly allows for groups based at Vanderbilt to include members from other universities, research institutes, or scholarly organizations based on mutual interest.

4.3. A DSpace community agrees to:

4.3.1. Define community membership and collection policies for its members;

4.3.2. Submit and describe content in accordance with Repository and metadata standards;

4.3.3. Notify the libraries of organizational changes that affect submissions;

4.3.4. Reply to a request for annual reconfirmation of Community information;

4.3.5. Observe University policies on DSpace and educate Community members regarding these policies;

4.3.6. Clear copyright for items submitted whose copyright owner is other than the author(s) or Vanderbilt;

4.3.7. Decide upon a submission workflow for each collection.

4.4. A DSpace Community retains the right to:

4.4.1. Establish policy regarding content to be submitted, within the DSpace guidelines;

4.4.2. Decide who may submit content within the Community;

4.4.3. Determine access to content at the item level either to Vanderbilt only, or unrestricted;

4.4.4. Receive a copy of submitted material on request;

4.4.5. Remove items and collections in accordance with the “Withdrawal Policy”;

4.4.6. Approve addition of or elimination of sub-communities;

4.4.7. Customize interfaces to Community content at the institutional repository level.

4.5. Internal Review, Certification, Access and Security

Once a DSpace community has been established (following guidelines described above) and a leader chosen, the University should largely withdraw from direct supervision. The community should express its own preference for review and certification standards and for access controls on their submissions. One can envision an array of appropriate practices depending on the purpose of the community and, as much as practical, the University should facilitate the choices made by the community.

4.6. Role of the Vanderbilt administration

The major roles of the administration with respect to DSpace are as follows:

4.6.1. Seek out and set up new DSpace communities as outlined above;

4.6.2. Determine if newly proposed user communities satisfy the criteria given above and authorize those that do;

4.6.3. Run and maintain the hardware and software of the Repository and pursue upgrades as needed;

4.6.4. Adjudicate internal and external conflicts concerning archived material and the identity of community leaders when required;

4.6.5. Assure the continuation of key communities, especially by verifying the existence of a community leader;

4.6.6. Maintain a loose oversight to ensure that the materials that communities archive are at least generally in compliance with the collection criteria.

5. Potential Uses of an Institutional Repository at Vanderbilt

This section addresses the services that could appropriately be provided by the Institutional Repository at Vanderbilt and what role the Library will play in delivering those services. The variety of document types described below clearly should be preserved, yet they currently are stored on Vanderbilt-supported servers, on faculty desktops, on departmental servers, and in VUSpace, where they are not accessible to other scholars and their long-term preservation is not being assured.

5.1. Role of the Library

The Library will provide guidance for the compatibility of formats used by communities with the approved formats that are supported by the Institutional Repository at Vanderbilt for long-term storage. The cost could be assumed by the educational programs of individual schools or by the research communities. The Library will provide the resources and technologies necessary to migrate and update the files stored here. The Library will assure the preservation of the files for unlimited time. The Library will issue a list of formats that it will support. This list must be updated as the technology changes or evolves.

5.1.1. *Teaching material* could be added at the request of the person who created it. Any material stored here should be compatible with OAK and an interface between DSpace and OAK should be a development goal.

5.1.2. *Research material* could be added at any time by the person who created it for long-term archival storage. Having an institutional repository for these materials would enable and reinforce interdisciplinary and inter-school collaborations throughout the campus.

5.1.3. *Research data* deemed worth storage for unlimited time by a community could be stored in the Institutional Repository at Vanderbilt.

5.1.4. *Working papers* could be stored for unlimited time, unless the individual who submitted any given paper requests its removal. If a working paper is removed, a record of its presence would be kept in DSpace.

5.1.5. *Reprints and preprint material* submitted for permanent retention and display as a manuscript should be reviewed before storage in the Institutional Repository at Vanderbilt. The community that proposes the manuscript should carry out the process of review. Once in the Institutional Repository at Vanderbilt, the manuscript would not be removed or modified unless forced to do so by legal issues (see Section 5.2 below). Addenda and corrections could be added to the manuscript, linked to the URL of the original, unmodified manuscript.

5.2. Copyright and Other Legal Issues

It is possible that some communities will place materials in their archive that are copyrighted, libelous, advocate illegal actions, etc. This is not a new issue. Current University websites have the same potential. The best course is to continue the current practice. DSpace communities are authorized by the University, and the community leader has control over what goes into the archive. Provided these are responsible people, legal issues should be rare. It is impractical in any event to filter DSpace content to ensure that no inappropriate material is included.

To address the copyright issue, when items are submitted to DSpace, authors grant to Vanderbilt University the non-exclusive rights to reproduce, distribute, and migrate their submissions as described in the "Non-Exclusive Distribution License". In this license, authors also represent that each submission is their original work, and that they either hold copyright, or have obtained permission of the copyright owner to grant Vanderbilt non-exclusive distribution rights.

When disputes over DSpace materials arise, the University will have to deal with them on a case-by-case basis. Should an item be reasonably challenged, it would most likely be removed from the repository. However, in the interest of scholarly communication, it would be a good thing for the University to resist any requirement by publishers that working paper versions of research be removed as a condition of publication. This is true independently of how DSpace is implemented.

Publishers are resolving some copyright-related issues at this time; for example, the publisher Elsevier is making the allowance that for their journals, "An author may post his version of the final paper on his personal web site and on his institution's web site (including its institutional repository)." The ability to quickly post a finished paper in the repository without hurting its potential for commercial publication will strengthen the usefulness of DSpace.

6. Hardware Requirements and Cost Projections

The DSpace software is designed to be flexible regarding hardware requirements. It is capable of running on equipment ranging from the very modest to servers costing $500,000 or more. However, current recommendations from the DSpace Federation, supported by reports from early adopters, are that a research university should strongly consider running DSpace initially on a dedicated intermediate class server with significant memory and disk storage.[4]

6.1 Near-term Hardware Requirements

We suggest following the recommendation of the DSpace Federation, hence to begin the repository pilot project with a SUN server system comparable to the following (approximate costs $30-$35K):

> SunFire 280R Server, two 900 MHz UltraSPARC-III Cu processors, 8MB Ecache, 2 GB
> memory, two 36GB 10,000 HH internal FCAL disk drives, DVD, 436 GB 10K RPM disks,
> SUN StorEdge A1000 rack-mountable with 1 HW RAID controller, 24MB std cache.

Since extensive experience with SUN equipment and the Solaris operating system exists on campus, both within the Library and within ITS, this should provide good performance and adequate storage for

[4] http://DSpace.org/faqs/index.html#hardware

the early phases of the Vanderbilt repository. The server should be housed in the controlled environment provided by the ITS network operations center.

6.2. Software and Content Services: Backup and Archiving

We will also need to provide for backup and archival services. For disaster-recovery purposes, we recommend that we contract with ITS to provide tape backup services using their existing equipment. In the short term, we would also need to store copies of the monthly backup tapes off-site, perhaps at the Library Annex, for archival purposes. Costs for storage media would be approximately $1200 per year. The additional costs associated with backup services provided by ITS and storage costs for the Library Annex would need to be developed. In addition to the direct hardware costs, we need to estimate the technical support costs associated with installation and maintenance of the hardware and installation and upgrades to the software. Extensive local customizations, enhancements to the software and user interface, and development of interfaces to systems outside of DSpace would require additional staff time. Since this will depend on community decisions and, to some extent, on the adoption rate, it is difficult to provide a firm estimate of these costs. The project will have other staffing needs that are not addressed here.

6.3. Intermediate Range Activities

In the intermediate term, as more items are stored on the DSpace server, storage may become a significant challenge. Rather than plan to expand server-attached storage indefinitely, we recommend that the Library consult with other departments or institutions to determine the feasibility of partnering with them for high-capacity storage services.

Currently, ITS has a scalable storage area network (SAN) with approximately 3 terabytes (TB) of available storage. Of this, all but .5 TB is committed to other projects (VUMail and VUSpace II). As DSpace storage needs increase, it will be advisable to consider collaborating with ITS to increase the available storage of their SAN to provide dedicated storage for DSpace. This would be particularly useful if some portion of VUSpace II could be integrated with DSpace in the future.

Should a partnership with ITS for high-capacity storage prove unfeasible, a second option would be to contract with OCLC for their Digital Archive service. This service is currently available for an annual fee.

6.4. Potential Outsourcing of Long Term Archival Storage

In the longer term, it will be important to contract with (or partner with) other institutions to develop and participate in digital preservation strategies and services. As mentioned above, OCLC provides a Digital Archive service for libraries and is one logical partner for this effort. It is likely that other potential partners will emerge as more universities develop institutional repositories.

7. Proposed Next Steps

7.1. Future Studies: Topics, Timelines

The Library is enthusiastic about taking responsibility for building an institutional repository at Vanderbilt. With the allocation of adequate resources to the repository, it could expand considerably because of the quantity of research done at Vanderbilt and the level of interest and need among faculty for digital storage that is more robust than current options. At other universities where institutional repositories are being built, the growth has begun slowly and then increased dramatically as the word spreads. The same model is anticipated for Vanderbilt’s implementation.

Existing Library staff is interested in the project and capable of performing the required functions. Reassigning them to new tasks related to the institutional repository could impact the Library’s ability to perform some less essential tasks, but the Library can probably initially manage the project by redefining existing positions. We are already acquiring a server and storage space for the near-term as already described. A rough timeline follows:

Phase 1 (Pilot Project):

(1) Receive approval of project policies by administration
(2) Complete prototype on test server, using a few communities as examples
(3) Resolve questions about migration, copyright, community structure, etc.
(4) Identify and train staff for long-term project
(5) Acquire server and storage

Phase 2:

(1) Begin production phase of project; migrate prototype documents to production server
(2) Pursue participation by additional faculty members
(3) Increase efforts to build communities and train faculty
(4) Adopt, develop, and implement enhancements as needed

Phase 3:

(1) Work with other campus technology implementations to investigate possible interfaces
(2) Continue growth of communities and collections
(3) Monitor growth and devote additional resources as needed

In Perpetuity: Continue growth, development, maintenance, management, and preservation

7.2. Pilot Project

For the pilot project, the Library plans to establish a few communities identified through discussions with the Institutional Repository Policy Committee. As the Library begins to build the repository and gain experience with it, staff will be able to establish procedures for adding various types of documents, accommodating features of each discipline and community. Many questions related to content, copyright, communities, and other issues will be identified and resolved. The repository will begin to grow slowly as the new equipment is acquired and installed and Library staff for the long-term project are identified and trained. Currently Vanderbilt has the DSpace software running on a test server. The three-member Library pilot project team consists of the Assistant University Librarian for Technical Services as project lead, the head of the Library Information Technology Services team, and our Music Cataloger, serving as metadata specialist.

7.3. Longer-Term Strategies for Creating and Investing in DSpace

As the project continues to grow, it will transition into a more long-term management plan. To instill confidence among those using the repository, the Library will need to ensure its quality, reliability and usefulness. The Library's commitment to developing faculty interest in DSpace will be an important determinant in whether and how it is used and what type of resource it becomes. We will persuade faculty groups to become involved, as the success of the repository depends on recognition of its value and usefulness, specific to faculty member and discipline, rather than any requirement to participate. To help the library build the repository, we recommend the appointment of a faculty Institutional Repository Advisory Board that will identify potential communities and contributors. This board will also advise on any other policy issues that might arise. It would replace the Institutional Repository Policy Committee once its charge is fulfilled.

An important role in developing the repository will be a librarian who visits various centers, departments, or groups of faculty and helps them determine how using DSpace could be of value to them. The Project Coordinator will take both this role and ongoing responsibility for managing the Library's internal activities related to the repository. Faculty, departments, and centers will be approached based on awareness of their interest through existing librarian-faculty liaisons, the Advisory Board, or the Project Coordinator.

Another Library staff role will be that of Trainer, to give each new DSpace community's members authorization to access their community and teach them to add documents. Communities will be created as requested for any group that satisfies the broad policy criteria described in section 4.2. These communities will work with the Project Coordinator and Trainer to establish their own standards for accuracy and appropriateness of documents or any other issues that arise concerning content. In addition to new documents and metadata added directly by faculty members, similar documents already in existence in other digital locations will gradually be migrated to bring collections together.

A Library Metadata Specialist will review and enhance metadata as necessary for consistency and easy access. This specialist will rely on the subject expertise of existing Library catalogers for questions about terminology appropriate to each field. Metadata review must not be allowed to backlog, as delays in this area could negatively impact the entire project. Therefore it may be necessary as faculty participation grows to increase staffing for metadata-related tasks.

As the repository grows, the Project Coordinator and Library Technology Staff will need to monitor storage needs and upgrade equipment. A very important preservation component will come into play as we master the skills of maintaining the documents in perpetuity and migrating through versions of software. The Preservation Librarian will become involved in the project at this level, mainly in an advisory capacity.

For the even longer term, the Library will be able to take advantage of enhancements to DSpace developed by its users worldwide. These enhancements may include interfaces with other applications, such as course management or peer review software, or e-portfolios, as well as improved user access through further developments in metadata structure and searching capabilities. We anticipate that ingestion software to automate the migration of existing documents and their metadata may also be on the horizon within the DSpace community.

7.4. Estimated Costs

Startup costs for the project are minimal, as the Library is using existing staff to develop the pilot project. A test server within the Library is adequate at this point for a small-scale prototype. To implement the near-term recommendations of the Committee, the Library has set in motion the purchase of the server and storage as recommended in section 6, using some year-end money.

As mentioned, if redefined existing positions can cover the project responsibilities, no added cost will be incurred at this time. In the long term, however, there is an opportunity cost in the loss of ability to perform other tasks, and a commitment to replace the reassigned staff at some point is desirable. Based on the description above, we estimate the following staffing will be needed to manage the project as the long-term phase begins:

(1) Project Coordinator (at least 1/4 FTE) to coordinate responsibilities within the libraries and develop communities among the faculty
(2) Trainer (1/4 FTE) to teach community members how to load documents
(3) Technology Staff (1/4 FTE) to implement and maintain the server and storage, upgrade software and hardware, preserve content, and develop software enhancements
(4) Metadata Specialist (at least 1/4 FTE to start) to review and enhance metadata; will rely on subject expertise of existing catalogers

Without knowing how quickly the repository will grow, it is impossible to know how long staffing at this level will be adequate. In addition, server and storage upgrades will require future funding. Several technology options have been mentioned; these options tend to become less expensive as time passes, and cost predictions now would not be accurate. If the project grows as we anticipate and it is properly supported and managed, its value to the University will easily justify its cost.

Institutional Repository Policy Committee 2003-2004

Richard Haglund, Chair
Department of Physics and Astronomy
Box 1807 Station B
2-7964
Richard.Haglund@vanderbilt.edu

Olivier Boutaud
Pharmacology/Medicine
Campus Zip 6602
514 RRB
3-7398
Olivier.boutaud@vanderbilt.edu

Jay Clayton
Department of English
Box 1654 Station B
2-2542
jay.clayton@vanderbilt.edu

David Cole
Department of Psychology and Human Development
Box 512 Peabody College
3-8712
david.cole@vanderbilt.edu

John Conley
Economics Department
Box 1819 Station B
2-2871
j.p.conley@vanderbilt.edu

Paul Dehart
Divinity School
411 21st Avenue South 37240
3-7516
p.dehart@vanderbilt.edu

Mona Frederick
Robert Penn Warren Center for the Humanities
Box 1534 Station B
3-6060
mona.Frederick@vanderbilt.edu

Doug Knight (on leave Spring 2004)
Divinity Dean's Office
234 Divinity School
3-5008
Douglas.a.knight@vanderbilt.edu

Tom Novak
Owen Graduate School of Management
Management Hall
2-3656
tom.novak@vanderbilt.edu

Tom Rasmussen
Law School
293A Law School Building
2-2810
Robert.Rasmussen@vanderbilt.edu

Janos Sztipanovits
Elecctrical Engineering and Computer Science Department
245 Hill Center, 372 Jacobs Hall
2-3455
janos.sztipanovits@vanderbilt.edu

Susan F. Wiltshire
Classical Studies Department
Box 18 Station B
3-4306
susan.f.Wiltshire@vanderbilt.edu

Ex-Officio Members

Michael Ames
Vanderbilt University Press
Box 1813 Sta B
2-3585
michael.ames@Vanderbilt.Edu

Paul Gherman
Office of the University Librarian
611 Gen Library Bldg
2-7120
paul.gherman@Vanderbilt.Edu

Annette Williams
Eskind Biomedical Library
campus zip 8340
6-3931
annette.williams@Vanderbilt.Edu

Roberta Winjum
Technical Services
Admin 700 Baker Bldg
3-3826
roberta.j.winjum@Vanderbilt.Edu

Technical Advisor

Jody Combs
Technology Team Leader
Library Information Technology Services
Office of the University Librarian
3-1229
joseph.d.combs@vanderbilt.edu