EDUCAUSE Center for Applied Research

Research Bulletin

Volume 2005, Issue 19

September 13, 2005

Digital Preservation in Action:

Toward a Campus-Wide Program

Richard Fyffe, University of Kansas

Deborah Ludwig, University of Kansas

Beth Forrest Warner, University of Kansas

4772 Walnut Street, Suite 206 ◆ Boulder, Colorado 80301-2538 ◆ www.educause.edu/ecar/

Overview

In an earlier EDUCAUSE Center for Applied Research (ECAR) research bulletin (Issue 18, 2005), we introduced a campus-wide perspective on digital preservation.[1] That bulletin, a companion to this one, outlined the stewardship responsibilities of the academy with respect to administrative content including student records, financial records, employment records, and so forth, as well as scholarly content such as research papers, experimental data sets, and course syllabi.

This bulletin explores a proposed model for establishing a digital preservation program in colleges and universities—requirements for educating the institutional community, developing roles and policies, and establishing an integrated technical architecture to support the complete life cycle of digital information. The model was developed at the University of Kansas as part of High-Velocity Change, High-Volume Collaboration (HVC[2]), an initiative that involved five work groups to explore collaborative learning spaces, digital preservation, and quality services for scholars, decision makers, and students.[2]

To provide a framework for discussion of the digital preservation program model, some background material from the earlier ECAR bulletin is included.

Highlights of Digital Preservation

The University of Kansas (KU) has proposed a holistic approach to digital preservation that begins upstream in the headwaters of digital creation. The philosophy driving the digital preservation concept will change our day-to-day management of information, both academic and administrative, in such a way that preservation becomes a natural outcome of normal practice. The focus of our approach is current digital assets; in this bulletin we do not address rescue of digital material in obsolete formats or on obsolete media or digitization of analog material.

Defining Digital Preservation

Digital preservation is the ongoing process of managing data for continuing access and use. It is an outcome of the organization's successful day-to-day management of its digital assets,[3] not the condition of an object. Effective digital preservation requires that we understand and attend to the full life cycle of digital objects. This life cycle has two key dimensions:

- *Chronological*—highlights the various stages of an object's life from the point of creation forward.
- *Functional*—highlights the different roles that are occupied through an object's interaction with creators, editors, users, stewards, policy makers, and others over the course of its lifetime.

No single step in an object's life cycle is the preservation step, and no single role is the preserving role. Rather, digital preservation is the whole complex of roles and operations designed around management of information for long-term accessibility and usability. However, we give particular emphasis to one role and one point in the life cycle, namely, the impact of choices made early in the object's life by creators (and other agents close to the creator) on the likelihood that it will remain accessible and usable over time. Those "upstream" choices can be exercised in four key areas:

- *Appraisal*—early recognition that the object will (or will not) be used in the future will encourage creators and stewards to make preservation-oriented choices.
- *Choice of file format*—some file formats are more likely to remain usable over long periods than others. In general, proprietary formats are less likely to be usable in the future than open formats.
- *Choice of storage location*—in general, files placed in networked storage devices maintained by professional administrators adhering to best practices are more likely to remain usable than files held on isolated disks (including workstation hard drives) if other preservation steps are followed.
- *Description (metadata)*—good documentation of the content of an asset and the circumstances under which it was created (name of creator, date, data source, file format, relationship to other components, and so forth) will enhance the likelihood that it can be discovered, rendered, and used effectively over time.

Developing a Digital Preservation Program

A successful institutional digital preservation program requires the interaction of three elements:

- *Education*—an educational program for faculty, staff, and administrators in the basic concepts and challenges in digital preservation and information management practices. The intent is to encourage good stewardship of institutional digital assets by creators and users of information objects and to encourage support for digital preservation by administrators and resource allocators.
- *Roles and policies*—a set of roles (functions exercised by staff) and institutional policies (to guide those roles) defined and assigned within the organization ensure that appropriate systems and services are implemented and maintained.
- *Integrated technical architecture*—designed around the complete life cycle of digital information, from creation forward. A technical infrastructure must be created that includes all the systems and services necessary to manage and access information over time.

Digital Preservation: Education

One key to the success of digital preservation planning on a university campus is the recruitment and involvement of staff throughout the institution. Content creators, policy

developers, and data stewards all have responsibilities for the preservation of digital materials that are generated as part of the university's mission.

An education program should be developed to provide the university community with the information needed to make appropriate decisions regarding preservation issues. The program might be structured around a series of workshops designed to provide basic information to promote a general awareness in addition to more advanced instruction for those whose specific job responsibilities include the ongoing maintenance of systems and data. The curriculum we have sketched includes five primary focus areas, shown in Figure 1 below.

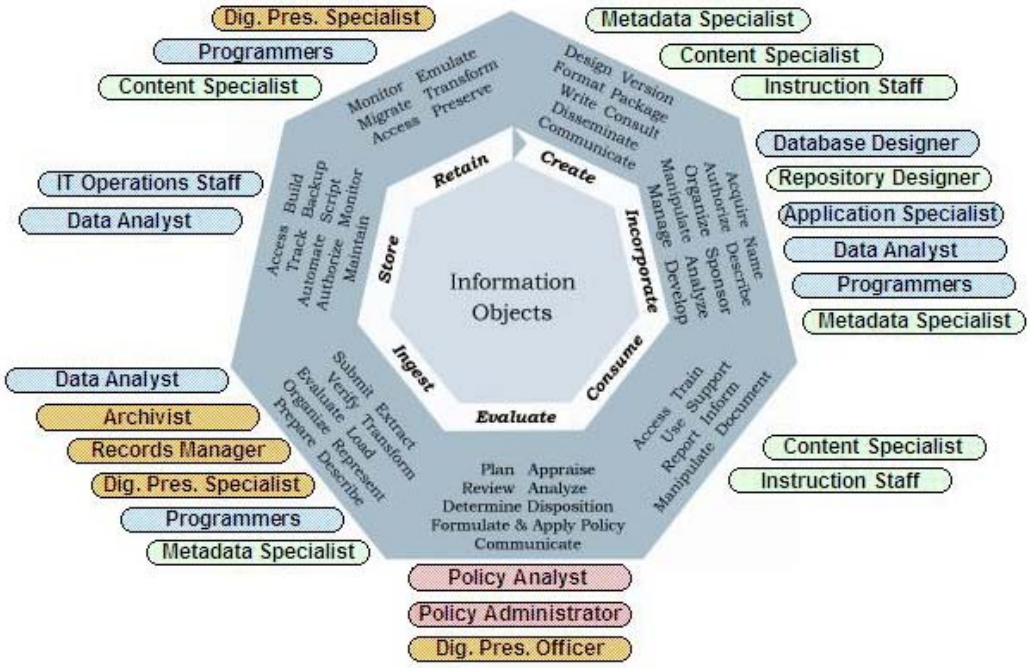
Figure 1. Sample Digital Preservation Curriculum Framework

1. General Awareness
   a. What is digital preservation and why is it important?
   b. Who is sponsoring digital preservation at the institution?
   c. Overview of roles
2. Information Life Cycle Management (ILM)
   a. The life cycle of information management
   b. Introduction to the OAIS model
   c. ILM Issues
   d. Stewardship issues
3. Information Storage Management and System Maintenance
   a. Archiving your existing data (what, where, how)
   b. Considerations for new systems
4. Best Practices and Standards
   a. Managing administrative data and electronic records (including e-mail)
   b. Creating digital collections
   c. Format standards
   d. Metadata
5. Legal Issues and University Policies
   a. Federal laws and regulations
   b. Records management
   c. Granting agency policies

Digital Preservation: Roles

From creation through retention, the organization's commitment to preservation is evident in the various processes that occur in the management of information objects and the roles assumed by individuals, systems, and services as illustrated in Figure 2.

Figure 2. Digital Preservation Roles



Roles are not necessarily synonymous with unique positions but instead indicate types of responsibility for processes. Suggested high-level processes and related roles include:

- *Creation*—brings ideas into existence as information objects representing data, texts, moving and still images, sound files, and other forms of usable expression across the spectrum of academic disciplines. The role of creator is shared by individuals, organizations, publishers, and information systems. An equally important role is the digital consultant or trainer who advises creators on the impact of various choices for long-term object viability. Moving digital preservation efforts upstream through early consultation is essential.

- *Incorporation*—brings information objects into the university environment through formal or informal acquisition and organization of information. While incorporation by itself does not guarantee digital preservation, it is an essential early step. Precise roles and processes may overlap with those required for object retention. Managers oversee aspects of incorporation, including policies, at a high level. Specialists have the knowledge and skills needed for a particular process to succeed and are also involved with services such as metadata modeling and development, data administration and stewardship, repository architecture and design, and quality assurance throughout the life cycle of information.

- *Consumption*—includes access, use, manipulation, and reuse of information. Persons, organizations, and systems that make information consumable and the

accompanying services that facilitate and control its use and reuse sustain the role of the consumer.

- *Evaluation*—considers the institutional framework of policy and standards that support the digital preservation program as decisions are made. Roles include legal counsel, data steward, preservation officer, records manager, and digital data policy development specialist.
- *Ingestion*—brings objects and their accompanying support systems into the retention phase of the life cycle. Ingestion may require further identification, organization, and transformation of information objects.
- *Storage*—information objects destined for permanent retention will require appropriate storage for the expected level of ongoing retrieval as well as ongoing backup and monitoring. Roles are similar to those currently required for storage of all university data and emphasize good management practice with carefully considered architectural requirements for long-term retention.
- *Retention*—as information objects age, the challenge increases in making them persistently available. Not only will the physical environment require ongoing monitoring, evaluation, and adjustment, but systems and services must exist to monitor future object availability and usability. Specialists knowledgeable about the university's digital assets, their information formats, and the requirements for making formats available to consumers as well as the processes required to migrate or emulate older data formats and systems are critical.

In addition, two other administrative roles are important to a successful digital preservation program:

- *Oversight*—although digital preservation should not be seen as a discrete activity, it is nonetheless important that someone in the organization be charged with remaining current with best practices, auditing local practices for consistency with established architecture and best practices, identifying needs for new resources, staff roles, and infrastructure, and keeping senior administration informed. This role could be assigned either to a single individual or to a small steering committee.
- *Sponsorship/advocacy*—especially in the early stages of a digital preservation program, it is important that a senior administrator with major authority for resource allocation and policy setting act as a sponsor and advocate for the program at the administrative level.

Digital Preservation: Policies

A successful digital preservation program is highly dependent on preservation-oriented practices being adopted within the university community. Faculty, administrators, and staff will need to know how to act and will need incentives to act in preservation-oriented ways. Therefore, a critical component of an effective preservation program is setting appropriate policies and guidelines including:

Appraisal: Not all university digital assets can be preserved. The following kinds of appraisal policies need to be developed to set priorities.

- *Administrative record appraisal and schedules*—many of the administrative records generated by the university can be designated in advance for discard, discard after a set term, or permanent retention. Some of these rules may be set by external regulation. Record schedules for the university should include electronic records.

- *Administrative data*—data generated by administrative transaction systems (specifically, ERPs—enterprise resource planning systems) should be appraised for retention in a data warehouse. Appropriate data structures, business rules, and retention rules should be defined and managed as part of the overall preservation program.

- *Nonadministrative digital assets*—this includes assets held by university units (including academic research products). Appraisal of digital assets not covered by administrative record schedules should be the responsibility of the custodial units, with guidance and consultation provided from a central authority. Universities should develop a template by which units can identify and inventory classes of digital assets under their stewardship and develop guidelines to help determine their relative priority for long-term retention.

Creation of New Content: Following best practices at the point of creation can help increase the potential for long-term preservation and usability of content. Areas to consider for development of policy and best practices include:

- *Recommended file formats*—a set of guidelines or recommendations should be developed for common preferred file formats (for word processing and spreadsheets, for example) and for designated specific support levels associated with them (for example, for certain formats the university would commit to preserving appearance and functionality, while for other formats it would commit only to preserving the bitstream).

- *Intellectual property for academic work by university authors and deposit of academic work in an institutional repository*—consideration should be given to a university or governance resolution calling upon all faculty to (a) deposit scholarly work of enduring value in their university's repository and (b) attempt to retain certain rights in their published scholarship, including the right to disseminate it through a university repository.

- *Permission to copy work on university servers for preservation administration*—in a digital environment, the preservation process is inherently one of copying—creating new copies of an original file onto new media or transforming files into new versions. Preservation of locally hosted academic work could be hampered by lack of clarity regarding university rights and responsibilities for work hosted on university equipment. Policy should be developed affirming the university's

expectation that preservation processes will be conducted on materials hosted on university systems.

Resources and Infrastructure: Policies outlining the requirements for institutional resources for managing digital assets over the long term are equally important.

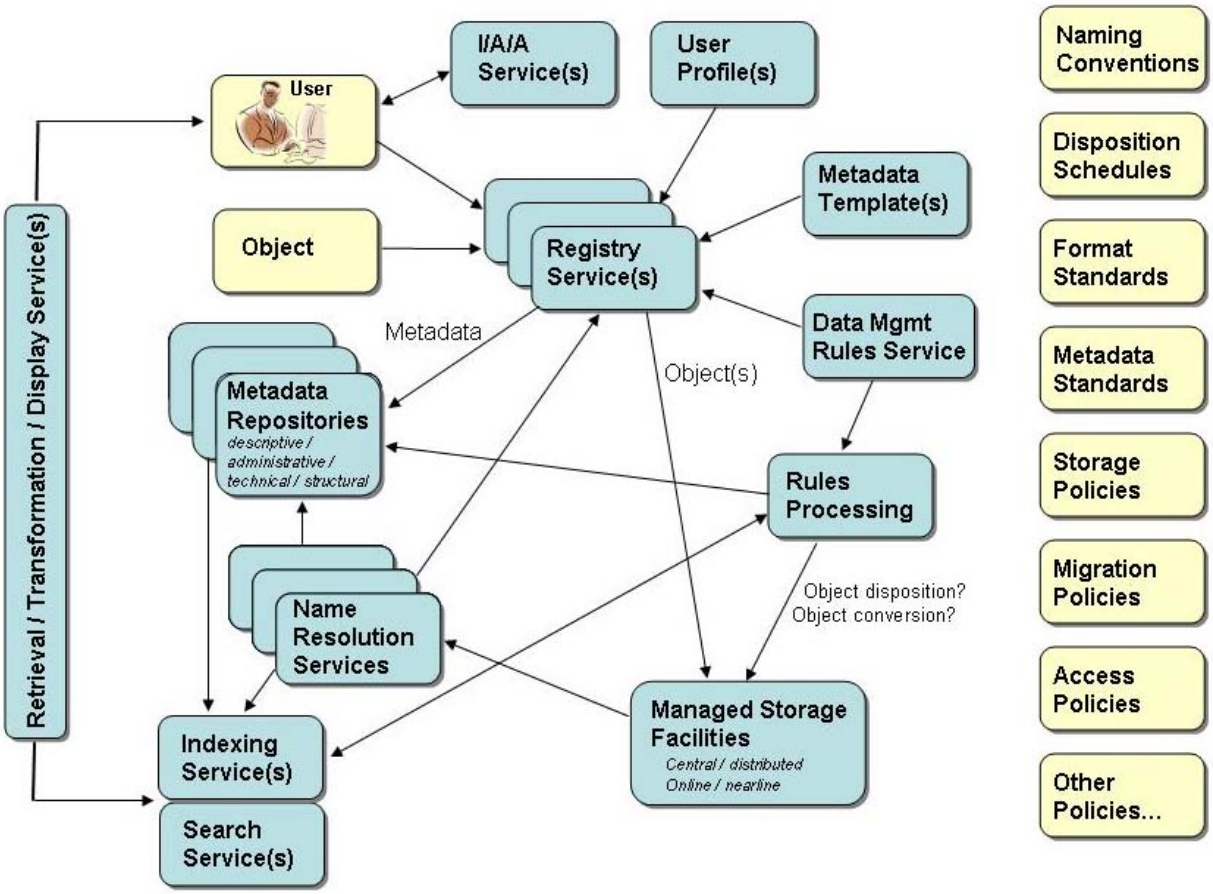
- *Best practices for noncentral technical support staff*—a set of guidelines should be developed to define best practice in the administration of servers and information processes regarding back-up protocols, security, and so forth.
- *Resource allocation*—a commitment to a university-wide digital preservation program cannot be undertaken without a clear understanding of funding sources. Some elements of the program will be funded centrally, while others may be the responsibility of individual units. Policy should be developed to identify those different elements and funding responsibilities.

Integrated Technical Architecture

Just as no one step in an object's life cycle is the preservation step and no one role is the preserving role, the technical infrastructure required to support digital object preservation cannot be defined separately from the basic technical infrastructure of the enterprise. Rather, a successful digital preservation program must be seen as part of the day-to-day operations and basic systems of the organization as much as possible in order to ensure long-term accessibility and usability.

The technical *infrastructure* should be defined within a consistent technical *architecture* for the enterprise. The purpose of the architecture is to guide the development of the overall information systems infrastructure. Within this architectural framework, a number of infrastructure components that contribute to life cycle information management can be identified as shown in Figure 3.

Figure 3. Digital Preservation Technical Infrastructure



While each of these components is part of an integrated approach, key aspects include:

- *Registry services*—provide the initial capture and registration of digital objects in the overall management infrastructure. These services draw on additional services such as user profiles, metadata templates, data management rules, and name resolution services to lead submitters through the processes and decisions required to collect information about the submitter/creator and the digital object, apply appropriate classification and disposition elements, determine the appropriate repositories, and assign the appropriate Uniform Resource Name (URN). These services should be available for interactive and batch submissions. They should be Web-based and/or include a desktop drop-box feature for ease of use.

- *Data management rules repository/service*—information required for both short- and long-term management of metadata and objects. These rules determine such issues as appropriate metadata templates/formats, acceptable object formats, classification/records groups, disposition of metadata and objects (methods and dates), access rights, and so forth. These rules are developed

based on approved standards, policies, and best practices and are associated with objects.

- *User profiles*—include basic information about the object submitter. Elements could include correct form of name, department affiliation(s), address, default repositories, default object classifications, default subject terms, default access rights designations, and so forth. Selected elements could be linked to authoritative databases such as directory services and authority files for name and address information, data management rules for current object classifications, and a list of available repositories.

- *Metadata repositories*—managed repositories for storing descriptive, administrative, and technical metadata. Managing digital files for ongoing accessibility will be highly dependent on preservation metadata. Metadata should be stored in a standardized format such as Metadata and Encoding Transmission Standard (METS). Metadata is maintained (or harvested) and indexed for searching and rules processing for long-term disposition and/or preservation management of objects.

- *Object repositories/managed storage facilities*—are managed repositories of digital objects. Repositories can be either centrally managed or distributed. Repositories should be registered and certified based on compliance with policies and procedures. Note: backup procedures and disaster recovery plans alone do not equate to a complete digital preservation plan and process.[4] File management procedures must tie into name resolution services to maintain accurate links to metadata repositories and records for objects.

- *Name resolution service(s)*—the persistence of links between resources and resource discovery services is essential to ensure long-term access to materials. A persistent identifier is a name for an object that will remain the same regardless of where the object is located (URN). When objects are moved, the current location must be associated with the persistent identifier through use of a resolver database. A resolver database is used to translate/map the name (URN) to a current location (Uniform Resource Locator, or URL).

- *Rules processing*—processes or utilities that can be invoked for a variety of management tasks such as applying record disposition rules, and so forth. Rules Processing must interact with the data management rules service, metadata repositories, file management utilities, conversion utilities, and name resolution services. Audit trails detailing actions taken should be maintained as part of the object's administrative/technical metadata.

- *Standards/policies/best practices*—the use of standards is one strategy used to assist in preserving the integrity of and access to digital information. Use of standards can also help ensure best practice in the management of digital information. Digital preservation policies give structure and general direction for specific actions.

As we noted in our previous research bulletin,

> Digital objects cannot simply be encapsulated and set aside; they are never permanently preserved like pickles in a jar…. Without a planning program for continued access to and usability of digital information, there can be no reasonable expectation that digital information created today will remain usable in a few years.[5]

Failure to develop a robust institutional digital preservation program is a failure of institutional accountability—accountability to donors, legislators, and taxpayers; to employees; to tuition-paying students and families; and to future generations of readers and users of a scholarly record hosted by higher-education institutions. Imagine being unable to assure access to course enrollment and student grades. Without this information valid transcripts could no longer be provided; student academic records could no longer be confirmed; information for financial aid requirements could not be reported; statistics vital to accreditation organizations could not be provided; and so forth. Eventually, accreditation for the institution could be lost as a result.

By contrast, implementation of a centralized system of object registration and management of the kind described in this Bulletin, beyond helping to fulfill an institution's mission, may also help to increase institutional productivity by facilitating the process of depositing and describing administrative data sets and reports, learning objects, and other materials likely to be needed in the future, and facilitating discovery of those objects when they are needed. Imagine an environment where, in the course of a morning's work at your desk, you can quickly and easily:

- Locate and review candidate vita and directly check publications, presentations, and reports online from institutional repositories.
- Analyze and model tuition data pulled from the institution's data warehouse.
- Write a reference for a former student after reviewing her e-portfolio and other academic information.
- Automatically reference check a journal paper, create metadata for it, and submit it to the publisher and the institutional repository via an electronic desktop dropbox.

More broadly, cultivating attention to institutional stewardship of digital assets opens opportunities for IT professionals, librarians, and archivists to play new roles throughout the organization, in technology development and implementation, policy development and review, and education. Moreover, their shared understanding of information life cycle can help the institution bridge organizational boundaries and integrate preservation perspectives more thoroughly into daily work. For example, the responsibility of preservation librarians—often restricted to the tangible assets held by just one campus agency—can naturally be extended to include user education, records management policy, and review of overall information policy. Similarly, subject librarians or

bibliographers, who are already familiar with the norms and practices of scholarly communication and publishing in specific disciplines and already engaged in liaison and instructional work, can play an important role in shifting our institutions from a preservation perspective largely focused on the published product of scholarship to one concerned with the researcher's whole practice of information management.

It is easy to see that the consequences of not adequately planning for long-term preservation of digital data point clearly to the need to treat digital information, from the point of its creation forward, as a core institutional asset with stringent maintenance requirements, rather than as an ancillary product of other core activities. It is also easy to see that while funding considerations for infrastructure, staff development, and education may initially seem high, the costs of failing to adequately address these issues will, in the long run, be unacceptable to the institution.

Key Questions to Ask

To assess their readiness to undertake a digital preservation program under the model proposed here, colleges and universities will want to ask the following key questions.

- Is there high-level commitment (president or provost level) to support digital preservation policies and efforts over the long term?
- Do we have a successful track record of collaboration and boundary-crossing among administrative and academic policy makers, technology units, and library units?
- What staffing and infrastructure components are already in place? Do staff have a clear understanding of component interactions and dependencies?
- What is our technical skill base and experience with identification/authentication/ authorization (I/A/A) services, repository services, metadata analysis, and system design?
- Do we have a plan in place to address these long-term issues and a mechanism for review and evaluation?

Where to Learn More

- B. Lavoie, *The Open Archival Information System Reference Model: Introductory Guide* (Dublin, Ohio: OCLC Online Computer Library Center and Digital Preservation Coalition, 2004), <http://www.dpconline.org/docs/lavoie_OAIS.pdf>.
- National Library of Australia, Preserving Access to Digital Information (PADI), <http://www.nla.gov.au/padi/index.html>. The PADI Web site is a subject gateway to digital preservation resources.

- *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access and the Research Libraries Group, <http://www.rlg.org/en/page.php?Page_ID=114>.
- RLG/OCLC Working Group on Digital Archive Attributes, *Trusted Digital Repositories: Attributes and Responsibilities* (Mountain View, Calif.: RLG, May 2002), <http://www.rlg.org/longterm/repositories.pdf>.

Endnotes

1. R. Fyffe, D. Ludwig, and B. F. Warner, "Digital Preservation: A Campus-Wide Perspective," (Boulder, Colo.: EDUCAUSE Center for Applied Research, Research Bulletin, Issue 18, 2005), <http://www.educause.edu/LibraryDetailPage/666?ID=ERB0518>.
2. High Velocity Change, High Volume Collaboration (HVC^2), University of Kansas, <http://www.ku.edu/~hvc2/>.
3. A digital asset is an electronic object that has value for some purpose. It may have been created digitally or it may have been digitized from a nondigital original source. Examples of digital assets include word processing documents, databases, Web sites, organizational records, digital recordings of musical performances, and so forth.

   To become part of a university's digital preservation program, the digital asset must support (directly or indirectly) the university's fundamental instructional, research, or public service missions. They can be academic or administrative in origin. A digital object should be considered a university digital asset if it satisfies one or more of the following criteria:

   - It was created in fulfillment of the research, teaching, or creative work of university faculty, staff, or students.
   - It is relevant to the planning, managing, operating, controlling, or auditing of administrative functions of an administrative or academic unit of the university.
   - It was purchased or licensed by the university in fulfillment of an academic or administrative function under a contract that permits continuing use of the asset (for example, certain electronic journals licensed through the library).

   Not all university digital assets will have equal priority for preservation. This working definition of *university digital asset* establishes minimum conditions for eligibility and does not determine any additional conditions that an object may need to meet to be included in the preservation program or to set priorities for preservation.
4. For additional details on management requirements, see RLG/OCLC Working Group on Digital Archive Attributes, *Trusted Digital Repositories: Attributes and Responsibilities* (Mountain View, Calif.: RLG, May 2002), <http://www.rlg.org/longterm/repositories.pdf> and the section on Storage at the PADI site, <http://www.nla.gov.au/padi/topics/10.html>.
5. Fyffe et al., op. cit., p. 2.

About the Authors

At the University of Kansas, Richard Fyffe (rfyffe@ku.edu) is Assistant Dean of Libraries for Scholarly Communication; Deborah Ludwig (dludwig@ku.edu) is Director, Enterprise Academic Systems; and Beth Forrest Warner (bwarner@ku.edu) is Assistant Vice Provost for Information Services (Strategic Initiatives).