Considerations for Campus Cyberinfrastructure Data Management Policy and Procedure Development

A Net@EDU Campus Cyberinfrastructure Working Group White Paper

October 2008
Abstract

Equally important as the physical cyberinfrastructure components designed to facilitate the transmission and storage of data are the policies and procedures for how that data is managed. The purpose of this document is not to dictate a standard set of policies colleges and universities should adhere to for managing data but rather to provide an overview of some of the many issues and challenges that campus policymakers should consider when undertaking this task. The authors’ aim is to help initiate a campus-wide dialogue on this topic.

Net@EDU helps higher education shape and take full advantage of emerging networking opportunities. Our activities span the spectrum of academic and cutting-edge research networking, from the administration of campus and state networks to identifying the most appropriate network projects for implementation to acting as a national center for research and advocacy in network policy issues. For more information, visit www.educause.edu/netatedu.
Introduction and General Comments

It may be impossible at this point to develop a model set of policies or procedures for data management in the campus cyberinfrastructure (CI) context; they are highly sensitive to local campus culture, assumptions, and requirements. Instead, we will outline issues and pitfalls authors of such campus policies and procedures need to consider.

Our strongest exhortation is for comprehensive and integrative thinking in developing these policies. No matter who leads the development of plans and policies for cyberinfrastructure (several reasonable scenarios exist), coherent CI policies must reflect and balance many different institutional concerns championed by various parts of the institutional leadership (for example, CIO, chief research officer, risk managers, faculty senate, deans and department heads, library leadership, general counsel) and also recognize that faculty work is situated in both campus and disciplinary contexts; as this work moves to a CI environment, both contexts must be maintained. In particular, the disciplinary context is typically interinstitutional in terms not only of actual research collaborations but also of sharing data and infrastructure, which brings with it problems of shared jurisdiction and federation—for example, issues surrounding security policy, audit, identity management, funding and accounting, institutional review boards and human subjects, and even national security regulations (in international settings). Both campus policy and actual campus CI implementation need to be sufficiently flexible to coexist sensibly with the disciplinary CI environments.

We also note that most policy work in this area to date has been high level; our sense is that the deeper the various connections and requirements, the more difficult policy formulation becomes. At the same time, overly general policy provides little guidance. Finding the right balance—and adjusting it as the base of experience and knowledge evolves—is a key challenge.

Finally, it is important that campus policymakers recognize that their work proceeds against a backdrop of important national developments, including the National Science Foundation DataNet program for data reuse and curation, the NSF Data Sharing Communities program, various National Institutes of Health data management and sharing initiatives (see, for example, the NIH National Center for Research Resources Strategic Plan 2009–2013), and the work of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Important and relevant work is occurring internationally; while a survey of these efforts is beyond this paper’s scope, note particularly the work in the United Kingdom funded by JISC, the Joint Information Systems Committee, in data curation (see the Digital Curation Center website at www.dcc.ac.uk), in disciplinary data services and institutional repositories, and in the feasibility study for a research data service (www.ukrds.ac.uk).
Considerations for Data Storage, Management, and Preservation

Backup
Institutions should differentiate backup, which is operational in nature, from curation, which is organized retention of specific data resources. They must be clear about the locus of responsibilities for backup, and where these responsibilities originate—responsibility for contracts and grants terms and conditions, overall responsibility for good research practice, responsibility for regulatory compliance requirements, employee responsibility to the organization, and so forth. The policy decisions for backup should be clearly articulated, followed, and audited.

What are the recovery time objectives (RTOs) and recovery point objectives (RPOs)? How are these met in the event of various forms of disasters? Who signs off on the appropriateness of backup policies? How are backup policies and procedures audited and evaluated? At what level in the organization are common backup requirements and standards set? How are backups financed, and who needs to pay? What are the confidentiality requirements of backups, and what data must be encrypted? How is the discoverability of information in backups handled? Is there information that should not be backed up? What happens to backups after the creator of the backup leaves the organization? Who has access to backups besides their creators? Should the people with media access also maintain access to the encryption keys necessary to access the data itself? How are media migrations handled? How is off-site storage handled? How are media sanitization and destruction handled?

Data Retention
Data retention applies to both data curation and backup data storage. How long does data need to be retained? Who decides, and who decides when data can be discarded? What criteria dictate the retention of data at an institution? What are the implicit and explicit requirements of institutional record-keeping policies? What are the implicit and explicit requirements of various funding bodies for research data? Who owns retained data, and who has access? Where does the responsibility for data retention lie, and how does it shift over time? How is data retention audited and evaluated?

Data Curation
Data curation is a set of activities that facilitate the long-term reuse of data, as well as its preservation to enable this reuse. Where does the responsibility for performing data curation lie at the various stages of the data life cycle? What is expected in terms of data curation? How are these expectations documented and communicated to data owners and users? How are questions about the accessibility and usability of data versus data integrity managed over time (for example, decisions about format migration)? Who is responsible for funding data curation? How is the responsibility and capability for data curation aggregated and distributed along disciplinary as opposed to institutional lines? In cases where there is no natural disciplinary center
for data curation, how can multiple institutions work together to achieve critical mass and economies of scale?

**Storage Provisioning**

How are the physical and logical storage and services implementations that address backup and preservation design requirements provisioned? Should the institution provide storage with multiple explicit levels of performance, redundancy, and backup (with presumably corollary variations in cost)? What storage designs are appropriate and proven to provide these explicit levels? Who finances, who maintains, and who provides user services for institutional storage environments? How can storage investments, both on campus and for redundant and geographically distributed and replicated storage, be pooled in order to achieve the best economies of scale? How should institutional strategies for storage provisioning be connected to disciplinary storage management strategies, particularly when there are performance considerations in the geographic placement of storage systems and their connectivity to various network, observational, and computational resources? A crucial question to consider is the potential role of institutional repository services as a possible framework for some of or all these functions.

**Data Sharing**

Many scholarly communities, and many funding agencies, are growing more interested in ensuring that scholarly data is available for long-term reuse. While this implies data preservation and data curation, it is important to recognize that the interests go beyond record-keeping to effectively advancing scholarship.

What are the data-sharing access mechanisms and services needed to advance scholarship? Are there common platforms that can interact with the storage-provisioning framework and address multidisciplinary needs? Should this access layer be provided in a discipline-specific manner? How does the access layer provide for data access and sharing?

What are the policies for data sharing? Where are these policies developed and vetted? Decisions include questions involving human subjects and institutional review boards, restricted information under various national security regimes, proprietary data, personally identifiable information, electronic personal health information, and requirements to affirmatively share data that funding agencies are imposing as conditions of contracts and grants. Where are the liabilities and responsibilities for data sharing and data security situated? What are the override mechanisms and procedures for gaining access to otherwise restricted data? How can responsibility for stewardship or access be transferred from one individual or group to another?

**Data Security and Identity Management**

Meaningful data sharing policies and practices cannot be developed except within the context of coherent security policy, identity management, and related authentication and authorization mechanisms.
Security policies must address and provide for the underlying business, compliance, and technical requirements of the system. They can be extremely complicated, if not impossible, in the face of contradictory user needs and regulatory mandates. Security policies must ensure:

- **Confidentiality**—the data or system is disclosed only as appropriate to authorized persons or organizations
- **Integrity**—the data or system is accurate, authentic, and complete
- **Availability**—the data or system is accessible when needed

Who is responsible for the security of the data, systems, and services? Who defines the security requirements for specific aspects of systems or data, and who ensures compliance with these requirements? How is information classified for varying security requirements? How is the value of the data versus the cost of security weighed and managed? What is the process for preparation and approvals of new or changed security policies?

We must be able to discuss identity in a common language in order to address who can do what. Identity management can and should be considered on an institutional basis and in the context of national and international collaborations, both disciplinary and cross-disciplinary. All these contexts are important for data sharing. What institutional identity management platforms exist? How is the integrity of those systems ensured (at a higher standard than the systems and data relying on their accuracy)? What broader identity management services are applicable to the campus and its users, and how do campus identity management systems federate or interoperate with them? How is trust ensured and verified with third-party services and federation providers, and what standards and legal agreements are applied to these technical implementations?

**Conclusions**

While this document does not offer a model set of policies or procedures for data management, it surveys many of the issues and pitfalls that authors of such campus policies and procedures should consider. Much work is being done nationally and internationally to address these issues and provide guidance. This complicated issue requires integrative, holistic thinking across an organization to develop appropriate policies to meet institutional needs. And, though an institution may not find a single existing policy document that will meet all its needs, addressing the questions posed here and developing the applicable institutional policies is a valuable endeavor that can reap long-term benefits.