The ECAR Data Management Working Group (ECAR-DM) has created a framework for research data storage as an aid for higher education institutions establishing and evaluating their institution’s research data storage efforts. (For more information, see the ECAR Working Groups website.) This paper describes areas for consideration and suggests graduated criteria to assist in that effort. The framework is built on six main categories:

1. Governance, Policy, and Oversight
2. Data Life Cycle
3. Use of Third-Party (Cloud) Providers of Storage
4. Communication and Collaboration
5. Infrastructure
6. Operation and Management

Each category then consists of several components for which the framework further provides maturity classifications: leading edge, industry standard, and lagging. ECAR-DM recognizes that each institution is unique in its requirements for research data storage and that an institution must choose its strategies and directions based on its individual needs. It is entirely possible for an institution to need to be leading edge in some areas, while lagging is appropriate in others. The framework presented here provides institutions with guidance for establishing new or enhancing existing institutional data storage strategies and plans. It should be noted that references throughout this document relating to federal policies are based only on current regulations and responses by institutions to those guidelines.

The framework operates with the understanding that research data storage planning and management starts with a conversation. In order to comprehend the storage needs of researchers, strategic communication and planning among institutional leadership and colleagues across departments and offices is critical to ascertain the support resources that are vital throughout all stages of the research data life cycle.

It is important to recognize that multiple perspectives all have a place at the table when institutions aim to develop and implement a comprehensive storage-management structure that works for the community as a whole.

Introduction

The framework developed and presented here is based on emerging trends and best practices at institutions of higher education across this country and around the world. Although a widely accepted foundation for actual standards relative to this topic does not exist at this time, the ECAR-DM provides this framework as a representation of the current environment existing across many of our campuses.
today. The framework and its components are based on several reliable sources for this type of information, including a review of current research and publications, along with a collection of common concerns and practices gleaned from institutions across the country, representing campuses of various sizes and levels of research activity.

It is not intended to represent a tool built on results of contemporary research findings, but rather to be an effort to reveal management practices at our institutions that will be helpful in developing emerging best practices. At the very least, the value of this framework for an institution lies in the conversations and discoveries that result when the community of stakeholders understands the important role of each component in the framework.

Representatives from throughout an institution bring unique and valuable skill sets that serve to strengthen the foundation of a model best suited to that community's needs.

Suggestions for Use

This document is intended to aid institutions as they develop the resources vital to sustaining the growing collection of research data being generated by their faculty, staff, and students. It emphasizes the critical role of institution-wide collaboration and strategic planning necessary for this development.

Institutions that use this framework to assess the current state of their research data environment will find themselves better positioned to diagnose problem areas and form strategic plans for improvement.

Some specific outcomes you can expect from this process include:

- Starting the conversation at your institution
  - Come to a common understanding of what research data storage maturity means across your institution.
  - Address accountability issues with stakeholders.

- Strategizing
  - Use or modify components of the matrix to identify where your institution is in the continuum of higher education institutions.
  - Identify specific areas in need of attention that, when addressed, will help your institution move up to the next level.
  - Use as leverage when justifying needs in planning and budgeting.

Understanding the Maturity Classifications

Rather than defaulting to a wide array of practices used by institutions to approach the challenges of supporting research data and its role as part of the larger picture of support for technology tools and resources across the institution, a more deliberate conversation intended to identify specific expectations can provide results that are both focused and targeted.

At the time of this writing, the trends in best practices of research data support tend to fall in three maturity levels:

**Leading Edge:** Ongoing communication, collaborative planning, and development—across the institution as well as vertically between administration and the institution—are integral to preparation
for deployment and sustainability of new technology resources that become available to support research data needs. Reviewing and piloting new tools and resources is a joint venture between those in academics, research, the libraries, and information technology. The resulting activities allow for growth and innovation specific to changing needs within each discipline.

**Industry Standard:** There may be a general awareness across the institution of the need to collaborate on strategic planning and implementation of a cohesive model of support, but developing this practice may be viewed as unnecessary to address randomly recurring issues. There tend to be hot spots or pockets of activity where collaboration occurs regularly between the central IT organization and distributed groups, such as the library. These incidents, although somewhat isolated, increase the likelihood of leveraging existing resources to maximize IT investments of staff, time, and dollars across a larger footprint of the institution.

**Lagging:** Any issues identified related to research data management are generally addressed in isolation or in an absence of any bigger picture of the institution’s goals for the role of research data in academics and research. There is typically minimal communication, either across departments or colleges or vertically between administration and the various departments with responsibilities that play a role in developing or supporting services. Planning for and adoption of new technology tools typically occur on an as-needed basis and/or after implementation of a new tool when support and maintenance issues become a priority.

As a result of the different roles that various components (identified below) play in research data storage, they may appear in multiple categories in the framework. As an example, the component “data sharing” may be addressed through an institution’s policies as it sets expectations of how a principal investigator (PI) coordinates with fellow researchers, and then that component may reappear as the institution determines protocols for authentication.

**Governance, Policy, and Oversight**

The goal of governance, policy, and oversight is to ensure alignment of IT services with institutional research objectives. It is essential that research data storage governance drive policy, funding, and research grant compliance. As Ricky Erway of OCLC Research notes:

> It is important to recognize the current uncertainty as to how data management support and services will be distributed among university, disciplinary, funder, and national and international stakeholders. In this complex environment, an institution must actively determine how it will manage and distribute data services internally. Various university players are important stakeholders in determining the appropriate governance structure to ensure efficient coordination; adequate security and regulatory compliance; and scalable, sustainable, and useful data management services to researchers.

Institutions must be keenly aware of the changing regulatory environment, particularly if they elect to deploy production infrastructure and/or services before these issues are settled.

**Research Data Storage Governance**

Research data storage governance comprises the IT leadership and advisory processes that determine needs and priorities to support research IT decision making. Governance specific to research data includes defining how decisions are made regarding funding, resource allocation, policy, and architecture.
Governance is defined in a formal, documented, and clearly understood framework for decision making, addressing all components of funding, resource allocation, policy, and architecture.

A formal governance process is documented but not well known by all involved.

The decision-making process is confusing and not clearly documented or shared.

### Policy

For the purposes of this framework, *policy* refers to the formal guidance needed to align an institution’s activities with external requirements or internal expectations. Policy describes the “what”—elements the institution relies on to achieve its institutional objectives—while procedures delve deeper into the operational “how” of these elements. Policy elements may include:

1. Definition of roles and responsibilities (PI, investigator, data custodian, research-based governing bodies, etc.)
2. Data definition and ownership
3. Data categories/classification
4. Data life-cycle issues
5. Data handling (based on the data classifications)
   a. Clarification of ownership of data (in general and for specific instances, e.g., if the PI leaves the institution)
   b. Access/availability of data (particularly before publication—release of data that contains sensitive information; Freedom of Information Act issues; etc.)
   c. Data sharing (proprietary, human subject, methods of data sharing, etc.)
   d. Transfer of research data
   e. Data destruction
6. Compliance
   a. Interactions with other institutional policies
   b. Relevant laws
   c. Grant rules
   d. Contracts
   e. Violations and sanctions

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Governance is defined in a formal, documented, and clearly understood framework for decision making, addressing all components of funding, resource allocation, policy, and architecture.</td>
<td>A formal governance process is documented but not well known by all involved.</td>
<td>The decision-making process is confusing and not clearly documented or shared.</td>
</tr>
<tr>
<td>Policies are frequently reviewed to ensure that they meet the objectives of the organization and accurately reflect institutional risk and compliance and operational needs.</td>
<td>Sanctioned policies may exist for specific areas, although such policies are infrequently reviewed for currency and accuracy.</td>
<td>No formal policy or practices are in place.</td>
</tr>
<tr>
<td>The community knows and understands their obligations under relevant policies.</td>
<td>Some parts of a policy are in place, and some practices are followed.</td>
<td>Any existing practices are not supported by a policy or the practice and the policy are not aligned.</td>
</tr>
<tr>
<td>Policy encourages intra- and interinstitutional data sharing, benefitting the researchers and the institution.</td>
<td>Policy related to data sharing exists in limited form.</td>
<td>No specific policy related to data sharing exists.</td>
</tr>
</tbody>
</table>
**Funding and Long-Term Sustainability**

There should be planning for long-term sustainability for research, including research initiated under grant-funded programs. Research data storage funding should be prioritized within the sustainability plans.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formal funding for storage is established for the life of the data.</td>
<td>Funding for data storage exists but is not formalized or prioritized.</td>
<td>No long-term plan for sustainability exists.</td>
</tr>
</tbody>
</table>

**Compliance with Research Grant Data Storage Requirements**

The data storage requirements of research grants, particularly federal grants, continue to increase, and these requirements must be addressed in each grant proposal and award.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>A formal process exists to understand and comply with storage requirements based on pending grant proposals and awards, as does a formal mechanism for the institution to respond to those requirements effectively and efficiently.</td>
<td>Compliance for research grant data storage is considered and addressed when new grants are awarded or new agency requirements are announced.</td>
<td>Compliance is addressed independently by grant PIs, often with little communication or engagement with centralized IT and other departments charged with maintaining the institution’s practices and policies.</td>
</tr>
</tbody>
</table>

**Data Life Cycle**

The data life cycle relates to the activities that involve research data. Included in this cycle are the creation, ingestion, processing, analysis, preservation, cataloging, authorizing access, and reuse of the data.

**Process**

Many processes are involved in the management of the data throughout its life cycle. These processes make up a comprehensive approach to managing an organization’s research data storage, involving procedures, policy, and practice as well as the underlying applications and technology. These processes utilize and move the data from one state of retention to others, based on the need at that point in the life cycle.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage processes and metadata are defined, executed, and maintained at an institutional level to ensure completeness and compliance across the institution.</td>
<td>Storage processes exist at varying levels within the institution. Execution of the storage processes occurs at those varying levels, with metadata being distributed across the institution.</td>
<td>Storage processes do not exist at an institutional level and are very limited at lower levels. Metadata are limited or nonexistent. Data are not managed institutionally, but rather from an individual researcher viewpoint.</td>
</tr>
</tbody>
</table>

**Archive**

This section deals with the practices and processes related to data that are no longer actively used in order to archive the data on a separate storage device or location for long-term retention. Consideration is given to indexing and search features for retrievability, whether driven by operational or compliance needs.
Leading Edge | Industry Standard | Lagging
---|---|---
The institution knows what data it possesses and where they are located, and it has the ability to access the data. There is a formal, designed process to update the data as needed to ensure archived data are accessible and retrievable. | Practices for archiving of data are inconsistent and not formalized. Some data may be protected and upgraded, and other data are not protected or upgraded. | There is no formal archive of data. PIs are not aware of community best practices for long-term archiving of data.

### Preservation and Curation

Preservation and curation includes the practices, processes, and activities performed to ensure that archived digital materials are accessible for as long as is required and/or useful. These activities are required to ensure that data in digital format are sustainable beyond the limits of media failure or technological change. The data may have been created as a result of daily business operations, material created in a digital format for a specific purpose (such as teaching), or created as the result of digitizing nondigital materials. The digital stores may be in the format that they were originally created or in a format-neutral state to allow access in the future. These activities also consider the need for deletion and/or destruction based on policy or law.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
</table>
The institution has long-term, sustainable preservation and curation architecture in place. | Modern processes and technology tools are in place for addressing preservation and curation of digital archives. | Preservation and curation technology is obsolete or aging. |
| Multisite replication and versioning of archived data is in practice. | The institution uses off-site replication for data. The archive only holds a single version of the archived data. | Institutionally, only a single archive site is available, if it exists at all. The archive holds only a single version of the archived data. |
| Content discovery exists to detect soon-to-be-obsolete formats. Best practice guidelines are in place for preservation, and preservation-amenable formats and preservation action plans are used. Ongoing technology oversight exists to proactively identify obsolescence. | Metadata exist in some form, but there is limited discovery of content. | There is limited, if any, metadata. Nonstandard archive formats are used. |
| A consultation practice is available. | Limited consultation is available to researchers. | Limited or nonexistent assistance is available for researchers. |

### Use of Third-Party (Cloud) Providers of Storage

Cloud service providers are offering for-fee, off-site storage. Some institutions will use these services for research data storage.

### Planning

Cloud service providers of storage are part of the overall storage planning.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
</table>
The role of third-party data storage services are planned and well understood as a part of the present and future research data storage strategy. | Third-party suppliers for research data storage are identified and considered but are not a formal part of the long-term strategy. | Each researcher is on his/her own to obtain third-party storage. |
Contracting

*Contracting* refers to the use of agreements with third-party providers of research data services.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>The institution centrally manages contract negotiation and compliance.</td>
<td>The institution centrally manages contract negotiation and compliance, but the vendor is not held to the same high standards as if the institution provided the services itself.</td>
<td>Contract negotiation and development with service providers is left to the PIs. Little to no guidance or oversight is provided by the institution to monitor the external service.</td>
</tr>
</tbody>
</table>

Agreements hold the vendor to the same high standards as if the institution provided the services itself, including availability, reliability, security, etc. Agreements ensure that researchers retain their intellectual property and that the data can be readily obtained should the PI no longer be available. Researchers are aware the agreement exists and know how to obtain the services if needed.

PIs are responsible for ensuring that agreements include processes for retaining intellectual property and access to the data if the PI separates from the institution.

PIs are responsible for ensuring that agreements include processes for retaining intellectual property and access to the data if the PI separates from the institution.

Communication and Collaboration

Internal organization and structure are needed for collaboration and interdepartmental planning for the day-to-day processes related to research data storage. By necessity (or by the nature of the resulting products and the skill sets and tools necessary to support them), these organizations include both library and IT departments. Additional participants such as sponsored programs and policy and guidance departments may also be needed. At the administrative level, collaboration and communication play a pivotal role in the ability of the institution to develop and oversee the framework for this research data storage service, ensuring the vision and process are shared across the community of users and supporting services. Additionally, as researchers from multiple institutions work on the same research project/data set, extramural communication and collaboration mechanisms are imperative. Finally, communication and collaboration activities are necessary to support the sharing of stored research data with the public.
Communication
This section refers to proactive communication regarding research data storage services to all stakeholders.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
</table>
| The institution is proactive in exploring research computing needs and reaching out to distributed IT departments, libraries, grants offices, and other stakeholders to partner on development and support. Environmental scans are used to understand user needs:  
- Identify research and academic requirements  
- Develop corresponding components within the research infrastructure  
- Continue communication by building relationships with the user community through a variety of communication channels: one-on-one communication with researchers, listening tours, brown-bag lunches, or special-interest groups focused on research-related technology activities  
- Communicate about research data storage options | Central and distributed IT departments respond to requests for assistance from researchers, focusing primarily on delivery of tools and resources that are already in place. Environmental scans are occasionally performed when there is sufficient demand to understand user needs. | Both central and distributed IT departments respond to requests from research and academics on an as-needed basis. No formal tool or process exists to understand user needs. |

Outreach efforts are long term. Awareness of current and future research needs emerge through one-on-one communication with colleagues across campus. There is no regular or formal communication process with the library or grants offices.

IT Department–Level Collaboration
The IT department collaborates and communicates with users of research data storage services to ensure optimal service delivery and enable future planning.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>The IT department leverages the research computing data storage services and infrastructure they provide to partner with researchers on grants.</td>
<td>The IT department is available to consult with researchers on grants as they identify data storage and infrastructure to support their research.</td>
<td>Requests to central or distributed IT departments for assistance in research data storage project development and grants are handled on a case-by-case basis or not done at all.</td>
</tr>
</tbody>
</table>

A fully funded research support point person or team is in place and tasked with managing project consultation, development, and implementation for research data storage services. A specific IT staff member or team is identified to address requests from researchers. No additional time allotment or salary is dedicated for this purpose. No specific point person is identified; requests are dealt with on an ad hoc basis.
Institution-Wide Communication and Collaboration

This section describes institution-wide collaboration and communication about current services and future needs for research data storage. There is a direct correlation between the level of collaboration between central IT and distributed IT shops and the availability of cross-purpose support. Cross-purpose tools and services (which can be used in both academic and research environments) provide IT with the ability to target technologies or skill sets that can help broad constituencies.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central IT and distributed IT shops collaborate consistently on developing new tools and providing resources.</td>
<td>Central IT and distributed IT units occasionally collaborate on developing new tools or providing resources.</td>
<td>Central IT and distributed IT units function independently on developing new tools or providing resources.</td>
</tr>
<tr>
<td>Central IT and distributed IT shops are consistently aware of existing services or support on campus; whether they are leveraged is a formal decision, not a reflection of awareness.</td>
<td>Central IT and distributed IT shops are generally aware of existing services or support on campus; opportunities to leverage may or may not be utilized, depending on awareness.</td>
<td>Central IT and distributed IT shops are likely to be unaware of existing services or support on campus, thus reducing the ability to leverage resources when the opportunity arises.</td>
</tr>
<tr>
<td>Potential of redundant services and support is unlikely.</td>
<td>Potential of redundant services and support is mitigated but may still exist.</td>
<td>Potential of redundant services and support is high, due to lack of awareness.</td>
</tr>
<tr>
<td>Results of consistent collaboration routinely create cross-purpose opportunities.</td>
<td>Occasional collaboration allows IT to assess needs and identify potential cross-purpose use. However, the results may be one-off cases where tools and/or services can only be used for a specific purpose or discipline.</td>
<td>The level of collaboration does not support identification of cross-purpose opportunities.</td>
</tr>
</tbody>
</table>

External Higher Education Collaboration

This section describes collaboration and communication with higher education colleagues and leaders on current and future research data storage and collaboration with entities outside the institution, including the government and private sector.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>The institution is deliberately aggressive in seeking out and forming partnerships with other state and regional institutions, consortia, external private industries, and government entities to maximize efficiencies and strengthen partnerships.</td>
<td>The institution occasionally forms partnerships with other higher education institutions, private industry, or government when convenient and when there is adequate time allotted for planning and deployment. Processes (such as formal partnership agreements and standard MOUs) are in place and renewed/reviewed regularly to ensure alignment with current vision and overall strategic plans.</td>
<td>The institution maintains a focus on serving the immediate campus community regardless of opportunities for potential partnerships with other state and regional institutions, external private industries, and government that could result in efficiencies in products and services. Central and distributed IT departments typically function independently unless requested or mandated by a specific initiative.</td>
</tr>
<tr>
<td>Managed centrally, the institution distributes IT support services across campus and across disciplines for the purpose of making cloud and externally provided services as economical as possible—or even free—providing the highest value to the customer and the institution.</td>
<td>Although not always centrally managed, some oversight is available to guide decisions on how and when resources should be designated for specific departments or projects, or distributed broadly across campus to maximize return on investment.</td>
<td>IT support services are most frequently designated for requests by specific client groups, projects, or departments, with little focus on application across more than one area or concern for long-term sustainability.</td>
</tr>
</tbody>
</table>
Finance Collaboration

Formal funding models for current and future research data storage services are developed and communicated.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
</table>
| Budgets, which are consistently developed with sustainability in mind:  
  • Include service fees  
  • Employ investment-based funding approach  
  • Exploit opportunities to leverage other (funding) external resources for maximum efficiencies and service value to the customer  
  • Consistently explore providing research data storage services to other institutions for a fee | Budgets are developed for required annual or biennial periods, with little emphasis given to long-term sustainability planning. | Budgets rely significantly on one-time (uncertain) funding sources and do not seek to share resources across units. Budgets have limited or no process to ensure funding sustainability. |

Infrastructure

Infrastructure is the hardware, software, and personnel expertise that form the lowest layer of the research data storage service stack. Examples include the physical plant of a data center (power, cooling, disaster hardening, etc.); the hardware required to store, protect, manage, and move data (disk arrays, tape backup systems, front-end servers for connection and authentication, high-performance networking, etc.); and the hardware- and service-level expertise (operators and system administrators) to keep it all working.

Data

This section focuses on data support, including whether tools for collection, extraction, reporting, analytics, modeling, etc., are available; whether training on tools exists; and whether metadata harvesting (the Open Archival Information System model) is in use. In addition, it includes whether access patterns are tracked to optimize storage usage and increase performance of the storage subsystems.\(^4\)
### Usability

**Leading Edge**
Support personnel can determine the type of data, its provenance, and its use case without accessing the data itself. Dashboard-type tools are available to track the size, structure, and access patterns of each data set. Storage managers use metadata to optimize the performance and cost of the institution’s data storage strategy.

**Industry Standard**
Metadata may be of varying organizational quality, and the data must be accessed manually.

**Lagging**
The condition of the metadata makes it difficult or impossible to monitor data contents or needs without accessing the data itself.

### Data Center

Known as the server farm or the computer room, the data center is where the majority of an enterprise’s servers, storage, and network equipment are located, operated, and managed. Critical attributes of a data center include adequate and reliable power, cooling, floor space, network connectivity, and monitoring.

<table>
<thead>
<tr>
<th>Plan for Growth</th>
<th><strong>Leading Edge</strong></th>
<th><strong>Industry Standard</strong></th>
<th><strong>Lagging</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The basic infrastructure for liquid-cooled machines (i.e., plumbing) is in place. Data center systems are designed with an eye to long-term growth. Environmental sustainability is a major factor in power systems design.</td>
<td>Floor space, power, and cooling are adequate for current systems and near-term growth.</td>
<td>Capacity issues are addressed in an ad hoc fashion.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power and Backup</th>
<th><strong>Leading Edge</strong></th>
<th><strong>Industry Standard</strong></th>
<th><strong>Lagging</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Electrical power is sufficient and reliable. UPS is available, sufficient, and properly maintained. Power generation is available, sufficient, and properly maintained. Appropriate management systems are in place to monitor the performance and efficiency of power systems and to enable predictive modeling for growth planning.</td>
<td>Power is clean; UPS is available.</td>
<td>Power is unreliable or backup generation unavailable. Cooling is inadequate to the combination of climate and internal heat load. Floor space is fully subscribed, or oversubscribed such that system density exceeds the cooling capacity.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Business Continuity/Disaster Planning</th>
<th><strong>Leading Edge</strong></th>
<th><strong>Industry Standard</strong></th>
<th><strong>Lagging</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>There is automatic failover to back-up centers, secondary data centers, outside hot sites, etc.</td>
<td>Defined procedures are in place for dealing with any system failures.</td>
<td>There is limited, if any, capacity to recover from severe outages. Backups may exist but are difficult to recover. There is redundancy in the hardware environment.</td>
</tr>
</tbody>
</table>

|                                        | Recovery procedures are exercised on a regular basis, including actual testing of failover. | Recovery procedures are reviewed and improved periodically and after any failure incident. Backups are available for recovery. | Recovery procedures are limited or nonexistent. |
**Network Infrastructure**

Network infrastructure includes the hardware and software resources of an entire network that enable network connectivity, communication, operations, and management of an enterprise network. Network infrastructure provides the communication path and services between users, processes, applications, services, and external networks/the Internet. This includes bandwidth internal to the storage systems (for instance, the connection between a data storage system and the backup systems that serve it) and connections to the outside world.

<table>
<thead>
<tr>
<th></th>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monitoring</strong></td>
<td>Automated monitoring systems provide comprehensive, real-time information to a 24/7/365 operations staff who respond following standard procedures whenever conditions outside predefined requirements and thresholds are reported. Advanced system integration improves the efficiency with which trained personnel monitor computing hardware and support systems.</td>
<td>Systems are monitored by professional staff, possibly remotely, 24 hours per day. Standard procedures for common error conditions are in place.</td>
<td>Monitoring is done unevenly or by underqualified personnel. Standard procedures for responding to error conditions are incomplete or not in place at all.</td>
</tr>
<tr>
<td><strong>Network Infrastructure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bandwidth</strong></td>
<td>A comprehensive plan is in place to grow bandwidth along with research traffic needs for the medium-to-long term.</td>
<td>Bandwidth to data storage systems is adequate to support current data flows and expected near-term growth.</td>
<td>Bandwidth to or within the data center is inadequate now or will be in the immediate future; the network is a bottleneck to data transfer.</td>
</tr>
<tr>
<td><strong>Bandwidth (Current)</strong></td>
<td>Predictive modeling exists, as does future planning that may include agreements with network providers that allow dynamic additions of bandwidth and pay for it on an as-used basis for specific research project requirements beyond typical needs. The planning includes both (flexible) preliminary hardware designs and administrative concerns such as funding.</td>
<td>There is limited network planning.</td>
<td>Plans to improve bandwidth are poor or incomplete.</td>
</tr>
<tr>
<td><strong>Traffic Management</strong></td>
<td>Traffic is managed via systematic automated monitoring with alerts, with automated failover.</td>
<td>Where connections from outside the data center are shared by other uses (commodity Internet, for instance), policies are in place to prioritize research traffic as needed.</td>
<td>Little monitoring or prioritization of traffic exists.</td>
</tr>
</tbody>
</table>

**Network Infrastructure**

Network infrastructure includes the hardware and software resources of an entire network that enable network connectivity, communication, operations, and management of an enterprise network. Network infrastructure provides the communication path and services between users, processes, applications, services, and external networks/the Internet. This includes bandwidth internal to the storage systems (for instance, the connection between a data storage system and the backup systems that serve it) and connections to the outside world.

<table>
<thead>
<tr>
<th></th>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monitoring</strong></td>
<td>Automated monitoring systems provide comprehensive, real-time information to a 24/7/365 operations staff who respond following standard procedures whenever conditions outside predefined requirements and thresholds are reported. Advanced system integration improves the efficiency with which trained personnel monitor computing hardware and support systems.</td>
<td>Systems are monitored by professional staff, possibly remotely, 24 hours per day. Standard procedures for common error conditions are in place.</td>
<td>Monitoring is done unevenly or by underqualified personnel. Standard procedures for responding to error conditions are incomplete or not in place at all.</td>
</tr>
<tr>
<td><strong>Network Infrastructure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bandwidth</strong></td>
<td>A comprehensive plan is in place to grow bandwidth along with research traffic needs for the medium-to-long term.</td>
<td>Bandwidth to data storage systems is adequate to support current data flows and expected near-term growth.</td>
<td>Bandwidth to or within the data center is inadequate now or will be in the immediate future; the network is a bottleneck to data transfer.</td>
</tr>
<tr>
<td><strong>Bandwidth (Current)</strong></td>
<td>Predictive modeling exists, as does future planning that may include agreements with network providers that allow dynamic additions of bandwidth and pay for it on an as-used basis for specific research project requirements beyond typical needs. The planning includes both (flexible) preliminary hardware designs and administrative concerns such as funding.</td>
<td>There is limited network planning.</td>
<td>Plans to improve bandwidth are poor or incomplete.</td>
</tr>
<tr>
<td><strong>Traffic Management</strong></td>
<td>Traffic is managed via systematic automated monitoring with alerts, with automated failover.</td>
<td>Where connections from outside the data center are shared by other uses (commodity Internet, for instance), policies are in place to prioritize research traffic as needed.</td>
<td>Little monitoring or prioritization of traffic exists.</td>
</tr>
</tbody>
</table>
## Identity Infrastructure

Identity infrastructure is a broad administrative area that deals with identifying individuals in a system and controlling their access to resources within that system by associating user rights and restrictions with the established identity.

<table>
<thead>
<tr>
<th></th>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plan</strong></td>
<td>Advanced identity management systems (including but not limited to enterprise directory, federated identity, and automated role-based/privilege-based authorization) have been fully implemented.</td>
<td>Some identity management is in place. Planning has begun to implement advanced features in the next three years.</td>
<td>Little or no planning or implementation has taken place for advanced identity management.</td>
</tr>
<tr>
<td><strong>Scope</strong></td>
<td>Identity management includes a fully functional federated ID and role-based privilege.</td>
<td>Identity management is consistent across institutional servers.</td>
<td>Identity management is not consistent across the institution.</td>
</tr>
<tr>
<td><strong>Procedures for Account Creation, Authentication, and Maintenance</strong></td>
<td>With new or changing roles, accounts are created and removed automatically. Federated identity processes are implemented.</td>
<td>Procedures for account creation and maintenance are quality controlled and rigorously followed. Defunct accounts are removed promptly.</td>
<td>Account creation and/or maintenance is error prone or labor intensive. Servers are inadequately secured, or authentication procedures are a security risk. Accounts are poorly tracked and/or not cleaned up when not in use.</td>
</tr>
<tr>
<td></td>
<td>Collaborative identity management practices are in place that allow for proactive intra- and interinstitutional data sharing.</td>
<td>Collaborative identity management practices are planned or in development that will allow data sharing to occur.</td>
<td>Without collaborative identity management practices, researchers are forced to find alternative means to share data.</td>
</tr>
<tr>
<td><strong>Access Policies</strong></td>
<td>Access controls are in place and policies and procedures are regularly audited. Tools are available to make this process simple.</td>
<td>Sensitive research data are identified and access is strictly controlled, in accordance with the particularly security requirements of the project and the data type.</td>
<td>Access procedures are inconsistent.</td>
</tr>
<tr>
<td></td>
<td>Advanced identification and authentication controls are implemented (such as multifactor authentication).</td>
<td>Password policies are strong, consistent, enforced, and reviewed regularly.</td>
<td>Password policies are inconsistent or nonexistent.</td>
</tr>
</tbody>
</table>
## Research Data Storage Architecture

This section addresses the storing of research data in a way that considers active use for research; storage capacity requirements; compliance with expectations of the institution and the funding entity (e.g., NSF or NIH); and archiving data that are no longer actively used to a separate data storage entity for long-term retention. Consideration is given to indexing and search features for retrievability, whether driven by operational or compliance needs.

<table>
<thead>
<tr>
<th></th>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Capacity</strong></td>
<td>Full capacity is planned for and capacity growth is automated or predictively planned. Agreements are in place for extra storage if needed. Adequate network bandwidth is in place to access the extra storage.</td>
<td>Capacity is adequate for the present and near future.</td>
<td>Data storage needs are met on an ad hoc basis as they arise. Storage is unstructured and/or data are poorly organized.</td>
</tr>
<tr>
<td><strong>Backup</strong></td>
<td>Backup systems are tailored to the research data set needs.</td>
<td>Backup systems are adequate for regular backup and timely restoration of data as needed.</td>
<td>Backup systems may be inadequate, leading to long intervals between backups or difficulty restoring lost data.</td>
</tr>
<tr>
<td><strong>Long-Term Strategy/Plan</strong></td>
<td>A long-term plan is in place to meet the institution’s data storage needs for the foreseeable future. Stored data sets are well structured and organized. Data storage hardware is expandable to meet new needs. A variety of storage types is offered to accommodate data from different sources (not just the institution’s data) and with different compliance requirements (HIPAA, FERPA, IP, Export Control, etc.). This phase is characterized by de-duplication and thin provisioning.</td>
<td>The institution has an adequate strategy to address performance (readily available, fast access to hot data), and capacity (slower access to infrequently accessed data or data that are retained for compliance purposes).</td>
<td>No or insufficient long-term planning is in place for the expected performance and capacity needs. Little or no consideration of ongoing compliance expectations of the institution or the funding entities is in place. No coordinated plan exists for long-term data life cycle management, including retention decisions, data format preservation, and archive.</td>
</tr>
<tr>
<td></td>
<td>Strategic architecture ensures that future data sharing requirements will be accommodated.</td>
<td>Long-term data sharing requirements have been considered and planning and budgeting requests have been made.</td>
<td>Little or no consideration and understanding exists of long-term data sharing requirements.</td>
</tr>
<tr>
<td></td>
<td>Architecture provides for integration with cloud-based storage vendors in a seamless manner for both central and distributed IT.</td>
<td>Integration with cloud storage providers is handled autonomously by distributed IT.</td>
<td>No coordinated integration with distributed campus storage or cloud storage is in place.</td>
</tr>
<tr>
<td></td>
<td>Architecture encompasses a federated approach to distributed campus storage.</td>
<td>Architecture acknowledges distributed campus storage, but integration is sporadic.</td>
<td>No formal architecture effort is in place.</td>
</tr>
</tbody>
</table>
Research Data Storage Connectivity Architecture

This area includes the components of campus research data architecture that focus on connectivity and storage system performance.

<table>
<thead>
<tr>
<th>Leading Edge</th>
<th>Industry Standard</th>
<th>Lagging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Connectivity</strong></td>
<td>Robust network connectivity is enabled. Network performance is regularly monitored.</td>
<td>Storage technologies (e.g., fiber channel) provide adequate connectivity for the current and near-term needs of both researchers and IT managers (i.e., it is possible to conduct a backup or restoration while under full user load without noticeable degradation in the user experience).</td>
</tr>
<tr>
<td>Tools, best practices, and techniques are in place to allow seamless sharing of data through easy-to-use interfaces and automated processes.</td>
<td>Tools are planned, in development, or in testing that allow seamless data sharing between researchers and other institutions.</td>
<td>Rudimentary tools are used to share data between researchers and other institutions.</td>
</tr>
<tr>
<td><strong>Storage System Performance</strong></td>
<td>Storage system performance is actively monitored via automated procedures. Alerts are sent. New, more efficient technologies are introduced as they become available. Long- and short-term planning are in place.</td>
<td>Storage system performance is adequate for the conditions of the medium-term future (~five years), or storage systems are designed to be improved incrementally to meet evolving needs. A variety of storage systems is available to meet varying use cases.</td>
</tr>
</tbody>
</table>

Operation and Management

The operation and management of data includes measuring the performance of data storage and access systems, responding to the needs and feedback from the data customers, supporting the use of data by researchers and PIs, and ongoing process improvement related to the data. Ancillary services, such as data visualization and high-performance computing, though related to data storage, are beyond the scope of this document.

Support for PIs

Support for PIs includes ongoing communication with and training for faculty, staff, and students in areas such as:

- Life-cycle storage and management processes, services, tools, and resources
- Data management plans
- Data-related changes in standards
- Changes to internal infrastructure
Ongoing training regarding research data storage offerings (e.g., policies, management, etc.) is available and advertised to the research community and required for PIs.

Some data management training is available to faculty, staff, and students.

Limited training is available. Infrequent communication of standards and infrastructure changes.

Elements of compliance are met, and opportunities to make operational improvements may be considered. Elements of compliance and operations are seamless.

Support for PI is based solely on compliance with campus and funding agency policies.

No assistance or guidance is available for development of data management plans. No formal structure exists to ensure PIs’ work is compliant or effective/efficient operationally.

Support for Individual Research Datasets

Formal oversight is in place to manage ingestion and preparation for later use and long-term management of the data. Data sets range from raw data to analyzed sets resulting from various stages of the research life cycle.

Data may consist of but are not limited to the following: data sets, publications, samples, physical collections, software, and models.

Services and processes are strongly coordinated.

Each data element has a single definition; clear metadata pertaining to the data and data set; processes in place to ensure data quality; data life-cycle management; processes in place for incident response; processes in place for appropriate authorization/authentication; and appropriate security tools in place.

Common types of services are available, but specialized services to address specific needs may not be.

Some services and processes are provided for research data management by either central or distributed IT, but there is little or no coordination between services. Partial support exists for dedicated staff, and funding is available to support these services and processes.

Researchers are on their own for managing their data sets. No services or processes are provided for research data management by either central or distributed IT. No dedicated staff or funding is identified to support these processes or services.

Customer Feedback, Process Improvement, and Performance Metrics

This area includes the employment of a formal assessment process to ensure accountability by those who provide support for the services and to the community of users. The focus of this component may include one or more of the following:

- An identified process for collaboration and communication in order to both obtain and share feedback on improvement and performance measures
- Defined evaluation strategies for the various phases of assessment, including collection of data, identification and measurement of results, and overall system assessment
## What’s Next

This document is but a first step in what we expect will be a long journey in the discussion specific to the world of research data management. Although higher education is quickly moving to address and support research data storage needs for academics, research, and administrative needs, the pace of data growth and the ways in which we are manipulating and accessing this data have thus far outpaced such efforts by many universities and colleges. By establishing some common language and descriptions to use when we describe our current state of support and where we hope to be headed, we are offering a way for institutions to understand generally where they stand in this space, to understand how they compare to other institutions (i.e., how they can enter into conversations with colleagues at other institutions to draw comparisons), and to identify where they may want to focus attention to make essential improvements.

But where next? We hope that this document will help expand this conversation and provide impetus for strategic discussion locally on your campus, with your peer institutions, and in higher education in general. The framework is intended to provide awareness of where your institution stands when it comes to supporting research data storage. It should be emphasized that this is not meant to be a one-time conversation, but rather an inroad into institution-wide coordination and awareness. The categories, components, and maturity classifications represented in this framework also allow different groups on a campus to use this document to focus on areas that they support so that the various groups can then come together for broader, coordinated conversations about deliberate planning and the process for addressing issues or weaknesses.

We expect that this framework will need to be reassessed in the future as we get a clearer picture of how to best support research data storage and as technologies (and our use of them) evolve. We also know that any guidelines and models we use will need to evolve to match new regulations by our granting agencies. In the meantime, it is also our hope that a broad use of this framework will enable and help focus key conversations across higher education and perhaps allow us to identify emerging trends—for instance, we may find widespread consensus that there is a broad lack of curatorial skill sets on our campuses or a need for liaison roles across departments, as a result of which we as a community will be able to work together to address these concerns in higher education in general.

We hope that you find this a useful document and look forward to future discussion.
Authors and Contributors

Special thanks go to the following ECAR Data Management Working Group (ECAR-DM) authors of and contributors to this report.

**Douglas Blair**  
Assistant Director, Administrative Computing  
Carnegie Mellon University

**Heather McCullough**  
Head, Digital Scholarship  
University of North Carolina Charlotte

**Barb Dawson**  
CIO, Health Information Technology  
Michigan State University

**Donald F. McMullen**  
Cyberinfrastructure Strategist  
Great Plains Network

**Michael Fary**  
Enterprise Data Architect  
University of Chicago

**Kim Owen, ECAR-DM Chair**  
Advanced Applications Outreach  
North Dakota State University

**Curtis W. Hillegas**  
Director of Research Computing  
Princeton University

**Mark Ratliff**  
Associate Director, Academic Technology Services and Digital Repository Architect  
Princeton University

**Brian W. Hopkins**  
Director, Mississippi Center for Supercomputing Research  
University of Mississippi

**Harry Williams**  
Chief Technology Officer  
Marist College

**Yolanda Lyons**  
Director of Internal Audit  
Colorado College

Notes

1. See, for example, the Australian National Data Service’s [Research Data Management Framework: Capability Maturity Guide](https://www.dataversity.net/aussie-ripley-on-research-data-management/), which defines “five levels of attainment or maturity which institutions may achieve in managing their research data,” from the initial level to the optimized level. From the United Kingdom, the [JISC Managing Research Data Project Maturity Model](https://www.jisc.ac.uk/infrastructure/managing-research-data) can be used as an “indicator of the institution’s current data management activity” (see also [The UWE case study](https://www.jisc.ac.uk/infrastructure/managing-research-data)). In the United States, efforts exist, but rarely at a national scale; for instance, Arden Kirkland, an MLIS student at Syracuse University’s iSchool, has published [A Capability Maturity Model for Research Data Management](https://www.syr.edu/programs/mlls/theses/2013/Kirkland.pdf), which presents a model that aims to provide “concrete guidance to analyze and assess the processes of RDM.”


3. For more information about the data life cycle, see the ACTI Data Management Working Group white paper, [Developing an Institutional Research Data Management Plan Service](https://www.acti.org/documents/2012-01/rdm_planervice.pdf). This document identifies common life cycle stages as “data creation (conceptualization), data collection and description, data storage, archiving and preservation, data access, discovery and analysis, and data reuse and transformation.”

4. Jerrold M. Grochow writes how analytics infrastructure is addressed in terms of three broad considerations: the processing cycle, the processing environment, and governance. This definition might be applied more broadly to IT infrastructure as data sources (operational systems, web, e-mail), tools (data warehouse, statistical analysis), processing environment (desktop, server, cloud), and governance (policies and procedures including but not limited to) decision making, security, and privacy matters. See [IT Infrastructure to Support Analytics: Laying the Groundwork for Institutional Analytics](https://www.ecar.org/infrastructure/to-support-analytics-laying-groundwork-institutional-analytics), research bulletin (Louisville, CO: ECAR, October 23, 2012).