WHITE PAPER
VERSION 1.0 FEB 2025

Research Data Management
Futureproofing

Dr Leslie D Almberg

Executive Summary

The intrinsic value of data is the cornerstone of research organisations, therefore data management techniques to preserve and optimise the return on investment in data must continuously evolve to keep pace with its rapid proliferation. Research data must be managed in ways that are sufficiently flexible to support the wide spectrum of research activities and facilitate new insights while pre-empting security threats and minimising costs. This document outlines the key considerations of research data management from multiple stakeholder perspectives and presents solutions designed to futureproof institutions' most invaluable asset.

Contents

Research Data Management (RDM) Challenges

Data is an elusive concept. The lifeblood of research and innovation; it can be ephemeral or timeless, irreproducible or a disposable commodity, structured or not, highly sensitive or a community resource, and terribly recalcitrant in the face of classification attempts. Like a single molecule of water in the ocean, a datum is all but lost once it runs through a researcher's fingers.

Assigning value to something intangible, ever evolving, and lacking a universal classification system presents challenges for institutions seeking to maximise their research investments. Data are only valuable when used, with reuse ostensibly boosting their value, yet there is no formula for predicting the lifetime flow-on effects of any data. Without a strategic approach, institutions struggle to balance data preservation, accessibility, and security to satisfy all stakeholders.

Calculating liabilities is more straightforward. The cost of managing research data extends far beyond providing storage. Institutions must invest in curation, migration, and infrastructure while addressing governance, compliance, and resiliency concerns. Additionally, data creators and holders face the risks associated with misuse, misinterpretation, and legal recourse once data are released into the wild, which disincentivises sharing practices.

As research data grow in volume, variety and velocity, standard management practices, which were highly dependent upon ad hoc, dispersed individual/lab/departmental efforts, fail. Universities and research institutions are actively transitioning to more secure, scalable, interoperable, and resilient infrastructure to meet evolving demand. The rush to adapt has, however, sometimes led to suboptimal solutions, such as cloud-based storage models with unpredictable retrieval costs and governance concerns. These challenges throw the prevailing wisdom of 'just keep everything' into question, forcing institutions to rethink long-term data storage management.

This section highlights the key challenges and frustrations facing data managers and researchers. The following section explores strategic scalable solutions designed to future-proof research data management.

Knowing what you have and where it is

Data buried is effectively data lost. Like the city of Pompeii smothered beneath the wrath of Mount Vesuvius, data loses meaning without context in the digital expanse between storage and retrieval. A time stamp and various, incomplete pieces of metadata provide little more than a record of deposition with limited value for extracting insights from stored data.

Extensive collections of laboratory or machine-generated data can quickly overwhelm storage systems with inconsistent structures, machine-readable-only file types, multiple versions, and machine-generated names. Naming conventions and file structure vary widely between individual researchers and teams, often making sense only to their creators. Without standardised, machine-actionable practices, institutional knowledge is frequently lost.

To maintain accessibility and integrity, research organisations require persistent policies and procedures for managing data holdings that facilitate search, curation, and governance from the moment of ingest. Setting permissions and decommission dates, tagging, and describing data are most effective when seamlessly integrated into the creation process. Retrospective classification is often inefficient and, without primary source knowledge, may be inaccurate.

Scaling existing systems and processes

Rapid improvements in the resolution and proliferation of digital cameras, microscopes, genomic sequencers, satellites, telescopes, seismometers, and MRI machines are a boon for scientific discovery, quality of human life, and disaster mitigation. Coupled with the explosion of AI-generated data and the growth of high-performance computing capacity to simulate climate systems, model neural connections, and map the universe, the data deluge has outpaced many existing institutional data management systems, necessitating scalable and adaptive solutions.

Managing, transferring and wrangling multiple copies and versions of enormous data sets is resource-intensive and costly. Many data archives lack efficient mechanisms to distinguish duplicates and original files, track active versus abandoned datasets, manage version histories, or automate retirement. Researchers often lack the training, time and motivation to develop and maintain disciplined data storage practices, creating difficulties for data managers down the line. Providing researchers with transparent, intuitive tools and workflows promotes integrating best practices seamlessly into their existing processes with minimal effort, making the entire curatorial process more efficient overall.

Managing access within and across institutions

Research endeavours are increasingly global and cross-disciplinary. This introduces a host of data management challenges that can outpace collaboration benefits. Without well-articulated management practices, procedures and tools, researchers may resort to the most convenient data sharing methods, potentially introducing vulnerabilities. Expedient approaches can pose security risks, strain institutional resources, and raise concerns regarding data provenance and trust.

Balancing data security with accessibility requires careful cost-benefit consideration. Secure data silos help protect assets but impede collaboration. Transferring copies wholesale breaks the chain of control and opens opportunities for corruption or invalidation. Best practice is sacrificed for timeliness or cooperation is deemed too difficult and avoided at the expense of better research outcomes.

Researchers seek to minimise impedances to their impact factor. Data managers want to minimise risks related to data loss or compromise and storage costs while maximising system efficiency to meet the needs of all users. Successful RDM requires tools that balance all stakeholder needs with institutional priorities, ensuring security, efficiency, and accessibility.

Deciding what to delete

Asking, 'Does it spark joy?' is not the recommended approach for culling ballooning data holdings. Some data are arguably valuable enough to warrant keeping indefinitely, while some have the potential to become valuable later, either alone or within a collection. However, the value of data is never initially clear; it can be dynamic or transient, complicating categorisation.

No one wants to be responsible for deleting something that is later deemed vital, yet how many petabytes of obsolete or orphaned data should languish to the end of time? What resources would be better allocated to decluttering data stores through de-duplication, retirement of the superseded, culling the corrupted, and sliding the coldest data onto ice? While data's greatest value may lie in unanticipated intersections, investment in such a trajectory is highly speculative.

Defining tiering, retirement and destruction policies to drive automated processes decouples these end-of-life directives from human error, indecision, and time constraints. Wrapping data within its own 'self-destruct button' upon creation, allows management to proceed with the lightest touch, and potentially no human intervention at all.

Creating archives of the future

Developing sustainable RDM policies to safeguard valuable data beyond the 'long now' raises a complex set of questions that researchers, data managers, institutional heads, funding bodies, and governments have grappled with for decades. Borgman's (2015) sixth provocation explores the challenge of aligning short-to-medium-term research funding cycles and shifting governmental priorities with the long-term effects of today's data investment decisions.

Stakeholders offer varying perspectives on what data should be preserved and why. What should be retained, for how long, and with what supporting documentation? Who bears the financial, curatorial, and management responsibilities and how are these transferred when organisations restructure, merge or dissolve? For whom should they be kept and how should access to read, write, or modify be granted?

Beyond these policy concerns, myriad technical considerations are essential. Bits rot. File formats become obsolete. Tape degrades. The energy demands of data centres contributes significant carbon emissions. The comedian Steven Wright once quipped, "You can't have everything, where would you put it?" His one-liner is a good reminder when thinking about data management, keeping everything isn't an option. A selective and strategic approach to data preservation is essential.

Amid these challenges, lead research institutions are overhauling their data management architecture with an eye to future requirements and technological developments. Princeton University, in particular, has made a bold commitment to building a 100-year data archive, which is explored in the case study at the end of this paper.

User Stories

A day in the life of the CIO/CTO

The role of chief information or technology officer in a research organisation varies widely as a function of institutional structure. They may need to be intensely inward focussed on the specifics of data governance and policies or externally tuned to trends and road mapping. Tensions between the microscopic and telescopic views, atop daily IT management requirements, leaves little room for strategic planning or driving institutional culture change.

There are many competing requirements to balance: Protecting sensitive data from inadvertent or malicious access without hindering permitted research activities; leveraging technologies to put data where researchers need it when they need it without blowing the budget; avoiding heavy dependence on one solution that could expose financial or security risks when disruptive technologies pop up or governance rules suddenly change. A CIO/CTO walks a metaphorical tightrope whilst juggling flames and spinning plates.

The proliferation of cloud-based data storage pushed many research institutions to reconsider their on-premises IT requirements, offloading the burden of maintaining and refreshing hardware. This was expected to significantly reduce expenses, but introduced new dilemmas around intellectual property, data provenance, and unpredictable cost structures, leaving many data officers scrambling for new solutions. Now the challenge is to find the best hybrids between on-prem and cloud storage and compute to optimise costs while adhering to regulations and facilitating low-impedance research activities. It's enough to leave data officers pulling out more than a few hairs.

Research organisations with significant data holdings recognise the intrinsic value of their data but often lack the resources and leadership for whole-scale institutional review and restructuring required to plan for the future. Being trapped in an endless game of whack-a-mole triaging the latest security threat, storage demand, or technological development leaves no opportunity to systematically overhaul data storage architecture, often leading to less-than-ideal, piecemeal solutions.

A day in the life of the research data storage administrator

In an overloaded/over-extended RDM system, the primary pressure on the storage administrator is to create space when things get snug. This is particularly difficult when collections of deeply nested folders lack clear logic or consistent naming/relationship conventions. While access control list (ACL) audits may reveal large chunks of old data long past last-access dates or created by researchers who have moved on, it tells the manager nothing about the data's value, or how many copies may be lurking elsewhere in the system.

Inevitably, records on this old data are sparse or non-existent, even within the research group that generated it. Unanswerable questions abound: Is it part of a current project or can it be archived to free up space? Does it have any governance requirements for retention at a particular security level or hard destruction past a given date? The default becomes to hold onto at least one copy of everything, just in case... Yet, as the cost of storage plateaus and concerns about security breaches and energy consumption mount, this ceases to be sustainable.

A day in the life of a researcher

Data is the driving force of research and the primary product of research endeavours. While researchers in data-intensive and big data fields are receiving more data management training during their early careers, RDM is not, nor should it be, their primary daily occupation. Furthermore, researchers in many domains and the latter stages of their careers more likely than not have had little to no RDM training, hamstringing their endeavours relative to those who have. All researchers wish to use their data optimally to answer their investigation questions with as much clarity and certainty as possible, share their results, then move onto the next phase of their research cycle. Time spent moving files between systems or digging through data archives is not considered time well spent.

Ideally, transforming and migrating data between formats and servers, governing and provisioning access, storing, backing up, archiving, and ultimately curating high value data is an invisible and effortless 'magic' that occurs in the background, freeing the researcher to focus on what they are trained for. In the perfect RDM world, data is easy and swift to locate and query, regardless of where it sits or who created it with what naming convention. It would be great to eliminate the need to manually clean up after every job by copying files to an archive, adhering to quota limits, moving data from HPC scratch to persistent or archive storage, etc. Tasks which are not only time consuming but often perplexing or opaque to non-IT specialists.

Automatic tagging with policy-based metadata upon creation and linking via other tools or AI to make them easier to find in the future would help researchers eke the most out of their data. Eliminating barriers to collegial data sharing, such as institutional permissions, file transfer size, version control, incompatible protocols and other logistical headaches can pave the way for increased research productivity and impact. Yet too few RDM environments support this array of functions, swallowing research hours with data wrangling, politics, and potential loss.

Comprehensive Research Data Management

Research data management is a broad, catch-all phrase for an ecosystem of tools, processes, roles and responsibilities to capture, store, serve, preserve, process, and curate research data. The different stakeholders across a research institution have diverse, and sometimes conflicting, needs, wants and propositional calculus. Implementing the 'least-worst' option, cobbled together from existing and 'affordable' pieces to grease the squeakiest wheels, stay within budget, and cover salient liabilities, is a common RDM approach. Adding piecemeal storage whenever capacity is reached, however, is not a sustainable data management strategy. Yet few institutions dedicate the necessary time and resources to develop a comprehensive strategic RDM plan.

This section explores several facets of RDM practices that can be reimagined to overcome the challenges set out in the previous section. Starting with a broad generic view, it provides key considerations to guide improved RDM practices. Heeding the pain points of end users and administrators guides the development of effective solutions at their intersection. Subsequently, we detail how Mediaflux's capabilities achieve these aims.

Search, find, use – metadata-powered software

Not all metadata is created equal. From basics of who, what, when and where, to detailing the why and how of every datum, the power of metadata is to harness and leverage empathy for the future self. Comprehensive metadata underpins any robust, resilient and sustainable data management strategy. To design a viable solution, however, the architect recognises that manually populating metadata for millions of files is prohibitive and must be automated. Not all data management solutions are created equal. Many software options are only able to supply and search basic file attribute metadata, not enabling tailored metadata population to make rediscovery easier and faster.

Beyond supplying system and file attribute metadata such as: file path, owner, creation, modification and last access times, plus read/write permissions, Mediaflux enables embedded and user-defined metadata including:

- a detailed description of the data in the file, with sufficient context for future use
- instrument specifications for machine-generated data, including:
  - instrument serial number
  - calibration date and references
  - standards used
  - operating conditions
- associated research lab and project file
- funding or cost centre codes (linked to any associated retention or sharing policies)
- principal investigator or contact, if different from creator
- required security protocols
- grant number – reference
- applicable governance protocols
- protection protocols
  - backup requirements
  - number of copies off site/on site
- archive/expiry/retirement/obsolesce/destroy dates or rules

Populating the full suite of operational, governance and descriptive metadata not only facilitates compliance, tiering, and deletion but also opens pathways for interdisciplinary collaboration and unanticipated future uses. To optimise the value of metadata, it must reside outside the data file as a sidecar file. Metadata files that exist as accessible, persistent, unified self-contained data repositories underpin powerful search and data intelligence capabilities.

Describe, tag, move, find, re-use and share

Partitioning metadata may seem counterintuitive, yet this practice yields a suite of invaluable options for both researchers and data managers. From rapid, ad hoc searching across the entirety of data, regardless of storage tier or location from any device, to facilitating discovery across all related data based on topic/domain/study/researcher/etc., stand-alone metadata are the key to overcoming many challenges facing RDM stakeholders. Furthermore, this provides a means to enhance security by only serving data about the data to verified researchers (without granting access to the files themselves) or by making entire projects effectively invisible to entities lacking appropriate permissions.

For end users, the ability to search by metadata saves precious research hours by eliminating the need to duplicate data across storage tiers or locations or request access simply to see what's out there. This capability not only accelerates finding one's own data but also discovering related or complimentary data sets. Comprehensive metadata bolsters the potential for re-use, facilitates data citation, and unlocks the potential for unanticipated cross-/interdisciplinary research outcomes. The aggregate of these benefits is a boon to research careers and society, boosting the overall return on data investment.

Providing researchers with a means to 'taste' their data, regardless of where it's stored, overcomes critical challenges faced by data managers. Giving researchers the confidence that they can quickly and easily find their data should the need arise in the future alleviates the dependency on expensive hot storage tiers, driving down costs and carbon emissions. It reduces unnecessary data migrations and cross-system duplications, freeing valuable storage space and bandwidth. Better still, this limits security risks through tight controls on the actual data without inhibiting knowledge creation.

Mediaflux achieves these requirements by providing extensive metadata harvesting, annotation, and cataloguing capabilities. As a comprehensive data and metadata management software platform, Mediaflux automatically indexes all types of data and metadata from any source or storage system. Automatically capturing the elements to keep discoverable and making them globally accessible accelerates search and discovery.

The cost and security benefits alone should convince any financial officer, vice chancellor or CEO that partitioning fully populated metadata files is worthwhile. Boosting research productivity, improving grant success, and amplifying the institutional reputations are added benefits.

Automated data decisioning

A good gardener knows weeding helps their garden thrive. So too, researchers know their labours can bear greater fruit when their data is easier to discover, yet dedicating time to culling and curating data within the constraints of grant and publishing time crunches can feel as worthwhile as plucking a few dandelions in the middle of a hailstorm. Hence, the responsibility for data migration, tiering, deduplication and retirement gets pushed back onto data managers, with curation offloaded to data librarians, if it's done at all.

The data manager, however, does not want to move or delete the wrong thing. Rather than risk being held culpable should

someone come looking for deleted files, research data managers may err on the side of retaining everything, even when institutional RDM plans, policies and procedures advise otherwise. This situation calls for machine-actionable, metadata-defined workflows and data decisioning. Placing the onus back on the researcher to input rules at the time of data generation and ingest frees data managers from future moral dilemmas. Moving decisioning to the time of data genesis, the researcher frees her future self from the tyranny of weeds like a gardener laying a thick bed of mulch before planting.

Scalable data decisioning must be metadata-triggered. Actions such as deduplicating and transferring data to appropriately secure colder storage tiers after an inactivity period or generating full reports of policy-defined destruction of all copies of sensitive data, can be streamlined via workflows to minimise human handling errors and time expenditure. This also relieves the burden from data managers to infer how valuable data is or decide what needs to be done with it.

Long-term curation

Maintaining data for the future presents a paradoxical challenge. Deciding what to delete, keep or curate, where, how, for how long and for whom introduces a mixed bag of value biases. Is old data being kept at the expense of generating new data? Which is more worthwhile? If you don't know what it is, can't find it, it costs too much to store, or exists in an unreadable obsolete machine-generated file format, is it a waste of resources to retain it?

While a complete exploration of the value proposition calculus is beyond the scope of this paper, we pose a few key considerations.

First, answers to the questions of long-term curation vary significantly by stakeholder and data type. The US National Science Board's 2005 Long-Lived Data report defines three categories of data for assigning preservation value: ***observational***, ***computational***, and ***experimental*** (in order of decreasing difficulty to replicate, therefore, decreasing retention necessity). Within each data category are nested subcategories, with greater or lesser preservation value based on objective (replicability) or subjective (potential impact) characteristics.

In conjunction with the broad data value classifications, researchers often confer additional significance to their data. Large, unique and unreproducible datasets are viewed with an eye to successive data reuse, recombination or mining opportunities. Regardless of intended future use, however, data managers observe the vast majority of data is deposited and never touched again. To incorporate both perspectives, RDM frameworks must facilitate streamlined reuse and trigger obsolescence responses. Associated metadata must contain sufficient context to guide use by other researchers if sharing is permitted, including governance, citation, and retention rules.

The effort invested in futureproofing data via comprehensive metadata population and management pays dividends in increasing appropriate future data use while reducing the load on data managers. Minimising the time spent on manual handling frees data managers to focus on edge cases and migrating pre-RDM data into new systems.

Breaking down data silos

Exponential data generation growth coupled with rapid technology development and ad hoc dispersed storage solutions has led institutions to recognise the concomitant risks and associated costs. Having intellectual assets physically spread across institutions/cities/states/nations/the globe frustrates effective management efforts, creates security risks, introduces unknown unknowns, and places barriers between potentially interrelated data.

Sequestering portions of an institution's data in distributed silos impedes, or makes impossible, search and discovery across the whole ecosystem. Needing to know where and how to look for what, or lacking access to even begin looking, diminishes the value of the stored information for the entire organisation.

Global namespaces

While there are valid and significant reasons to store data across physically distanced spaces and different media, it can construct barriers to search and discovery. This may oblige data managers to dedicate undue time and resources helping researchers recall and access their own data. Researchers are likely to end up with multiple versions stored in various places for ease of access and backup. All stakeholders benefit from a system that minimises these inefficiencies.

Introducing a metadata-aware global namespace creates a universal directory of all data, indifferent to structure, type or location, to appear as if available within a single repository. With unified control seen through a single pane of glass, scanning and accessing unstructured data across the entire data ecosystem of object storage buckets, object stores, and shares becomes seamless. Allowing users to fluidly locate and interact with data wherever it is physically stored, simplifies access and management, thereby enhancing the user experience.

When coupled with comprehensive metadata, a global-namespace-controlled storage architecture gives a data manager the omniscience required to orchestrate storage optimisation. In concert with this, researchers' efforts are optimised, with time spent searching for and transferring data minimised. Facilitating search and discovery, based on any parameter, across the entirety of data holding, regardless of location, the metadata-aware global namespace capability of Mediaflux expands the realm of research possibilities.

Workflow automation, not data wrangling

From raw, machine-generated output to published interpretations, data traverses diverse complex landscapes of processes. Multifaceted teams of technicians, students, researchers, data scientists, and principal investigators may need to pass the baton along an intricate information relay, merge thousands of disparate datasets, cull missing or corrupt data points or files, divert portions of datasets into different streams, or perform any number of operations specific to their field and investigation. In some domains, this data wrangling can consume upwards of 80 percent of a researchers' time, leaving them anxious to get on with things further down the pipeline.

Given the onerous and time-prohibitive nature of manually populating metadata to fully curate data, it is little wonder time-poor researchers resist adding this upstream task. When RDM workflows are framed as a means to make "research data understandable and procedures reproducible for people unfamiliar with the work," (Borycz, 2021) there is little incentive for the researcher to dedicate energy to metadata population beyond intrinsic altruism or the desire to avoid institutional penalties. However, reframing the value proposition as reducing data wrangling time, thereby freeing the researcher to do more downstream work, makes adopting RDM practices more attractive.

Automated workflows can leverage metadata scraping to streamline documentation. User- and machine-generated metadata, such as project and sample descriptions, cost centre, instrument model, calibration details, and operating conditions, can be combined with automatically extracted file information (creation time, size, type, etc.) and applied to all files generated. The linked data and metadata can then be transferred through the analysis pipeline together, with different access controls for each.

Employing automated workflows means the researcher or technician need only complete one 'sticky note' to attach to all relevant data. At each subsequent step of the analysis process, more notes can be 'stuck on' to keep a complete record of all events occurring in a dataset's lifecycle. Each additional 'sticky note' boosts the data's value by making it more verifiable and versatile, with additional potential for novel or unanticipated future reuse.

Mediaflux encompasses a full suite of data orchestration tools, which support both simple and complex workflows, delivering data to external applications and systems for further downstream processing. It can be used to move data directly from instruments or sensors through machine-actionable pre- and post-processing steps to meet a wide range of research data management requirements. The inherent flexibility of Mediaflux allows it to be tailored from small, limited use cases and easily scaled to support hundreds of billions of files.

Building security and trust

Structured filesystem attributes allow only very coarse data classifications, such as age, ownership and permissions. Lacking sufficient data about the data within research environments, the default is to be highly restrictive with all data to adhere to certain governance requirements for a particular subset. For example, this may occur when a portion of a lab's data is subject to HIPAA restrictions. If access for specific file information must be limited, the data manager may elect to play it safe and treat all data from that lab, or even the entire system on which it's stored, as HIPAA-regulated data. This certainly reduces the risk of failing HIPAA compliance protocols; however, it may also unnecessarily restrict access to non-sensitive data, potentially obstructing research outcomes.

Further to sensitive research data procedural risk, data repositories are vulnerable to a variety of internal and external threats to their sovereignty and security. Research organisations need to consider how their storage solutions are engineered to prevent loss or data corruption, not just recover from it. Creating cyber resilience requires minimising the potential attack surface of a system, placing multifactor authentication protocols in the data path, using strong encryption, memory-resilient programming languages, and approval workflows, among other strategic steps.

Data governance

Controlling who can take what actions, when, with what information and under what circumstances, using what methods is the key function of data governance. Research institutions are held to a wide range of governance requirements contingent on their location, partners, and type of research activities. From state to federal, or even international privacy laws protecting human subject data and personal information, to sensitive, temporal geospatial information, juggling access rights and deletion/retention policies is mired in multiply dependent variables, which are subject to change.

Organisations require a means to control not just access to, but visibility and permissions to read, write or delete sensitive data. They need regulation-triggered workflows to automate tiering to compliant cold storage and prevent movement into uncontrolled cloud backup. All actions performed on data must leave an audit trail to ensure adequate records of required data destruction. Comprehensive metadata can assist in driving automated workflows to minimise institutional risk exposure.

The actor-role model integral to Mediaflux facilitates seamless compliance by controlling who can view, read, write or destroy data based on the role they are assigned, not where the data sits. This means there is no need to move data or lock down projects with personnel changes. Mediaflux also has the capability to automate tiering and policy-driven decisioning to ensure data is retained as long as required and destroyed as necessary with a complete audit record.

Data provenance

Documenting an unambiguous chain of data custody bolsters the veracity of research results while enabling other researchers to replicate analyses or challenge outcomes. Tracing the origins, custody, and ownership of research data is a means of holding data creators accountable for their work. Further, it creates the conditions for tracking how other researchers repurpose that data in subsequent investigations.

Implementing file versioning is essential for delivering high-resolution data provenance records. Locking the initial instance forces each successive file modification to generate a new copy and audit record, capturing all versions of that file. A fine-grained audit trail reveals who changed what, when, and how, leaving open no questions of responsibility.

Beyond versioning, tracing provenance can even serve the researcher as they progress into the publication phase. A complete data handling history containing all metadata changes provides much fundamental information required for a research paper. The entire research process can be documented as a file moves through a workflow and quickly siphoned off into the tedious, but vital, methods section. Mediaflux transparently provisions this functionality.

Multifactor authentication and authorisation

Research institutions are treasure troves of data. The value of these keystone assets justifies deploying every reasonable protection to ensure their integrity. Research organisations are vulnerable both to external threats (e.g. data being held to ransom, unauthorised persons accessing and leveraging sensitive human subject or defence data, or compromising data integrity) and internal risks, both deliberate and accidental, such as data manipulation, leaks, or unintended deletion.

Multifactor authentication (MFA) is now common practice for accessing all sensitive data. From email to finances, end-users expect to provide more than just a username and password to access any account. Requiring something additional the user knows/is/has is still only a perimeter barrier, within which data may be vulnerable to malicious actors who manage to breach the last line of security.

Placing MFA in the data path builds a final line of defence against data loss or system compromise. Triggering MFA when an actor requests access to highly sensitive data or to perform critical actions such as migrating, coping or deleting significant data holdings and requiring confirmation by a second or third person eliminates many potential pathways for data sabotage. Accidental deletion becomes impossible and other malicious internal activities are eliminated along multiple checkpoints.

Case Studies

University of Melbourne

The University of Melbourne (UoM) was an early visionary in overhauling their institution-wide RDM infrastructure and practices. The introduction of an Australian national institutional RDM framework and Research Data Storage Infrastructure (RDSI) motivated the university sector to adopt new policies, procedures, infrastructure and services in an effort to improve data resilience and RDM coordination within and between institutions. As part of this initiative, a team of UoM stakeholders investigated the facility, transfer speed and volume, and end-to-end management pipeline capabilities of several RDM tools to meet the university's requirements. The agility of Mediaflux and RDSI's endorsement made it the logical solution for addressing the university's mandate to become equipped to tackle Big Data challenges.

Well into the second decade of this successful collaboration, Mediaflux continues to be a key component in UoM's research data management solution as an overarching storage gateway, and for storage tiering orchestration and disaster mitigation. Arcitecta built their award-winning Data Mover application in response to UoM's requests and needs, now moving 1-1.5 PB annually. The university has also recently made the decision to implement Mediaflux MFA in the data path for all users to enhance security protocols.

The successful institutional culture-change endeavour hinged on several critical decisions made at the inception. Starting from a fundamental review of HPC capacity and compute-intensity, change leaders were able to make a solid case for a multi-million-dollar investment based on what they had and where they needed to go with it. Deans were persuaded to lend support to an even larger investment than requested upon illuminating how many full-time employees were simply managing desktop storage, and by the case for inclusive, generalised and specialised shared infrastructure to improve efficiency, security, human resourcing, and resilience. They also accepted the necessity for flexibility and support for differing requirements and uptake barriers.

This long process of change ultimately amplified research outcomes and efficiency, improved collaboration, streamlined data management and enhanced data security, access and sustainability, securing the university's position as the top research university in Australia and a top-tier higher education institution globally.

Princeton University

Like UoM, Princeton has centrally funded the Tiger Data system to increase the capacity, reliability, functionality, and performance across the institution. The mission hinges on the ability to describe, tag, seamlessly move, easily find, re-use, and share all research data with collaborators. Their vision encapsulates a three-component system: a custom user interface for researchers and faculty to interact with their data, supported by Mediaflux middleware, built upon heterogeneous, heat-map-based tiered storage. It is designed to support all different types and lifecycles of data through metadata management and a user-friendly front end.

The goal of Tiger Data is to store data more efficiently, promote culture change around RDM across the institution (including researchers in less computationally intense humanities fields), and forge mutually beneficial partnerships between the library, data services, and enterprise IT. They are currently developing policies and processes for effective and efficient automated data tiering and workflows.

References

Borgman, C.L. (2015) 'Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press DOI:https://doi.org/10.7551/mitpress/9963.001.0001.

Borycz, J. (2021) 'Implementing Data Management Workflows in Research Groups Through Integrated Library Consultancy', Data Science Journal, 20(1), p. 9. DOI:https://doi.org/10.5334/dsj-2021-009.

National Science Board (2005) *Long-Lived Data Collections*. http://www.nsf.gov/pubs/2005/nsb0540.