ARCITECTA
OPERATING SYSTEMS FOR META+DATA

G-Know

Next generation research data management

Abstract

Organizations understand the value of research, but sometimes lack the right data management techniques to preserve and unlock that value. Ultimately, organizations need to manage research data in ways that are flexible enough to yield rich insights, minimize costs, and also comply with data security, governance and protection requirements. This document provides a technical overview of these issues, and suggests that a simple phased approach to data management can avoid many of the traditional pitfalls and roadblocks.

Acknowledgments

This paper was produced by:

Authors: **Ladwig, Greg**
Beasley, Graham

Arcitecta
talk@arcitecta.com
+1 720 583 4006
www.arcitecta.com

G-Know
info@g-know.com
858.356.5790
www.g-know.com

The information in this publication is provided “as is.” The authors and the organizations they work for make no representations or warranties of any kind concerning to the information in this publication.

Use, copying, and distribution requires written permission from the authors.

Executive Summary

When it comes to research data management, organizations don't always know where to start, because they don't know which data will prove most relevant and valuable in future. Their impulse is to be conservative and keep everything – but then the challenge becomes how to find it later. Another problem is keeping people in the loop: staff turnover often means that the importance of the data, and where it resides, becomes 'Tribal Knowledge'.

This document will outline the features and processes that enable organizations to get the most from their research data – maximizing its value, minimizing costs, while addressing data security, governance and protection issues. The principles outlined here will emphasize life science companies and research institutes, but can apply equally to any organization with growing collections of disparate data. It's better to start somewhere than nowhere – and in the authors' experience, a phased approach to data management is often the most successful.

The major challenges enterprises face when managing data

The research community faces three major challenges:

- **Finding your data:** We don't know what data we have now, where it is, or its provenance. We also don't know who will need what data in the future. Wouldn't it be great if there was some magical Rosetta Stone that helped us decode, classify and manage all this disparate data? While there is no all-in-one solution, a combination of data curation, search and governance can make data easy to find. In particular, metadata may turn out to be as useful to the research community as it was revolutionary to the music industry.

- **Data Explosion:** The resolution of digital cameras, microscopes, genomic sequencers and MRI machines is increasing – but storing the data these devices capture is not getting cheaper at the same rate. Compression algorithms do get better, and storage of all types is going down in price, but not enough to keep pace with improved resolution.

- **Workflow Automation:** To process complex datasets, the research community often relies on high-performance computing (HPC) clusters, which could be located on premises, at a partner's site or in the cloud (e.g. Elastic Computing Cloud). Researchers must marshal data from their own research facility into an input job that an HPC resource can process. Then they must retrieve and store the result data set, annotating everything so the experiment can be duplicated or modified in the future.

It is all about the data

These three challenges – finding, storing and processing data – are all about how to extract the most value from data. And while all organizations must handle the usual business files, research institutes and life science companies shoulder an additional burden – their huge collections of laboratory or machine-generated data. Many research organizations take a laissez-faire approach to data management . They give their users a home directory for their own files, plus shared storage for all the lab-generated data, but will often leave the structure and management of that data to the user's discretion.

But machine data presents a unique data management challenge, because unlike many business documents, these files aren't easy for human users to read, identify and organize They typically aren't in 'industry standard' file types, and the data in them is machine-readable only. Lacking a directory, 'Readme', 'sidecar' or header files, a user can't determine exactly what's in the files, and faces an arduous task to get value from them.

Take a common example of a machine data file – a music .mp3 file. If someone gave you an mp3 of your favorite artist, you would likely recognize the artist, and maybe that specific song, but you couldn't determine much more – when and where it was performed, the other artists, and so on.

Now, what if someone offered you a 10TB royalty-free music collection? Sounds great, right? But it comes with a few caveats: all the file names are randomly generated, the artist and classification info is blank, and all the files are stored in a single, very large file folder. How valuable is that data? For most people, not at all, because it's an almost insurmountable task to get to the files that have value, without sorting through the whole folder.

Many life science companies are in a similar position with their data holdings. They have very large volumes of machine data, with machine-generated names, and often the only information about that data is the

filesystem Access Control List (ACL), which shows the data's owner, the date it was written and the last access date. Typically, this kind of data is stored within a very deep folder structure that lacks any clear meaning. In short, the company has no overall understanding of its own data collection. Instead, it relies on the individuals who generated the data to interpret the limited parts of that data they can access, and to remember later what it was –and even then, only as long as they still work at the company.

To make matters worse, many research organizations use a regular rotation of graduate students as temporary employees. So, in a data management scenario that relies heavily on staff continuity, these institutions may have significantly more former employees than current employees, and therefore their large volumes of data have effectively been orphaned, because the organization lacks any primary source of information about that data.

The only way to ensure this data retains its value is to adopt information-collecting processes about the data, and methods to store and query this critical information. Without these meta-processes, companies will end up with vast stores of data, about which they know very little. This reduces the organization's data to little more than a historical archive – it proves that work was done in the past, but too often it yields no valuable insights now or in the future.

Other data management challenges

Data security and governance

As the volume of data increases and the storage environment grows, organizations face evolving requirements to secure, govern and understand their data. Data privacy laws such as GDPR and HIPPA change over time and new laws governing data privacy and retention policies are continually being introduced.

Total data volume & the 'five V's'

Many in the data management field talk about their 'five Vs': Volume, Velocity, Variety, Veracity and Value. Volume and Velocity refer to the rapid increase in the amount and pace of data as it streams from more devices, at continually improving resolutions. The growth in DNA sequencing alone is estimated to double every 7 months based on historical growth rates from the early 2000s to 2025 and beyond[1]. This velocity requires enough fast storage to write and process the data when an organization receives it; and it forces storage capacity to expand at a rate commensurate with growth. Taken together, these factors require a data management strategy as well as increased budget and administrative resources to expand systems and migrate data.

Variety, meanwhile, is endemic in life sciences data, which spans everything from well structured electronic health records to unstructured videos or machine generated data. Veracity (sometimes called Variability) refers to the 'trustworthiness' of the data: there is little point in saving and managing 'big data' if it comes from a nebulous or uncertain source. For example, if a microscope was not calibrated, does it make sense to save the several hundred terabytes of data that microscope produced over ten years? Do you want to base your analyses on this data? Finally, Value is one of this paper's guiding points. If it takes you longer than two years to find insights from the data being collected, or it costs too much to store, or it's in a machine-generated file with an obsolete file type, then what really is its value? Is it worth the organizational resources to retain the data at all?

Enabling comprehensive search

With this increase in data comes an increase in the number of storage devices, repositories, and locations. The result is data silos, which makes searching for information across the entire organization very challenging, or in some cases impossible. Not knowing where and what to look for, or a lack of access to data reduces the value to the organization of the information being stored.

[1]Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical? PLoS Biol 13(7): e1002195.

User Stories

A day in the life of the CIO/CTO

Being a CIO or CTO in the research sciences can be challenging, but dealing with privacy laws and the fear of data leakage or piracy can be overwhelming. It's your impulse to play defense – to build a fort around your data. But on the other hand, enabling data reuse and sharing is often significantly more productive and profitable – especially when a disruptive technology or set of rules comes along and challenges your status quo.

Cloud-based data storage and processing has changed the game so much that many questioned whether on-premises IT infrastructure still made sense at all. But it's risky to outsource your intellectual property (IP) – the lifeblood of your company. If you aren't organizing your own data holdings, and you want to recall and reuse a small portion of them, how would you find the needed data? And if data ever goes missing or costs too much to access regularly, then how would that affect research?

Perhaps it's better to maintain and manage the data on-premises, and use the cloud only for elastic processing. You'll also need to consider 'sunk cost' issues, asking yourself: "Didn't we just purchase the latest, fastest storage with a five-year service contract only last year?" Then there are all those privacy laws and patent lawyers questioning whether we invented this... The list goes on, and the pendulum of opportunities and challenges will always swing. The one certainty is change.

Large organizations often view appropriate information usage as a company asset. Indeed, many firms deal specifically in IP, data, or both – they are essentially in the licensing business. For these organizations, safeguarding the data is paramount; so, besides their CIO and CTO positions they have 'C-level' positions that deal specifically with data – for example, a "Chief Data Officer" is typically responsible for enterprise-wide governance , while a "Chief Knowledge Officer" is responsible for ensuring the organization maximizes the value it achieves through stewardship of 'knowledge'. Other data-centric positions can include a "Chief Data Officer & Chief Data Scientist", "Chief Digital Officer", and "Chief Digital Transformation Officer".

A day in the life of the IT storage administrator

Essentially, a storage administrator's job is to find room inside a storage unit that has become so full that soon it won't be able to perform new writes. And it's a tough job if you look at the file structure and can't see any clear or obvious logic within its byzantine collection of deep, deep folders. You review the ACLs, searching for a large chunk of old data you could remove, and find a few large directories from a few years ago, created by several different people who no longer work at the company.

So you reach out to these departed employees' former managers, trying to find out what the data is – is it part of a current project, or can it be archived to free up space? But the managers aren't much help – when they review the data, they, too, only see deep folders of mysterious machine data. They know their former colleagues were involved in a few projects – some of which are ongoing, others that have ended – but they can't tell what data belongs to which project. Lacking a better way to resolve the issue quickly, together the storage administrator and the managers draw a straw and archive a set of data, hoping for the best.

It's an unfortunately familiar scenario – and it's usually only a matter of time before those involved rue the decision they made.

A day in the life of a researcher

Perhaps the most important person of all in a research data management study is the user – the researcher whose work generates the data. As the researcher, you want to do the job you trained for – managing files across different systems, or finding things that have been archived, is not your idea of time well spent at the office. IT isn't your field, and often you know little about it.

In an ideal environment, a researcher could quickly find backups or old versions of files. You could do 'smart searches' of relevant data, whether you created that data yourself or received it from colleagues at a geographically remote site. It would be great if you didn't have to manually clean up after yourself – copying files to archive; respecting quota limits; moving data from an HPC job to persistent or archive storage, and so on. It would also be great if files could be automatically tagged with metadata as they were created, based on an easily understandable policy, and if they were linked with other tools or artificial intelligence (AI) to make them easier to find in the future. This would free you to share data with colleagues without spending lots of time dealing with permissions or copying complicated addresses, dealing with file transfer protocols (FTP), hypertext transfer protocols (HTTP), and other logistical headaches.

Unfortunately, for many researchers much of the above is a pipe dream. Their current reality is spending way too much time – in their opinion – "wrangling" data from one place to another, and dealing with the office politics of data, such as irate bosses telling them they're putting too much data on the "expensive, fast storage".

A more comprehensive data management strategy

It's clear that we badly need a more comprehensive

data management strategy – one that goes beyond adding yet another storage unit when existing capacity is (yet again) exhausted. We need a strategy that not only manages storage costs, but also maintains and extends the data's utility and value, and includes data security, governance and protection as an integral part of the strategy.

Where we've come from: a cost-only based approach

One of the earliest approaches to data management was to treat it as a cost-based problem based on infrastructure limitations. Because it was defined as an infrastructure-only problem, the solution was similarly defined around infrastructure – in this case, a data archive. The infrastructure team looked at the data, using the information available about that data to archive a portion of that data to cheaper storage – typically based upon its last access date.

While new technology has updated this approach – data tiering solutions with more robust tiers, with a public or private cloud as the final repository – this has not changed the view that data is an infrastructure challenge. As a result, better, faster and cheaper data storage has come at a cost of limiting the value of the information, and has made security, governance and protection increasingly difficult.

This is not to say that storage tiering and archive is the wrong answer. Instead, it's an incomplete solution. Effective data management requires looking beyond infrastructure. We must redefine the problem to find a more comprehensive, robust and useful solution. Until now, the problem has been defined in terms of reducing data storage costs. Instead, the problem should be: "How do we maximize the value of our data at the lowest cost?" Tiering or archiving will certainly play a role – but we will also require a far more detailed understanding of our data than file names, ACLs and directories alone can provide. To extract the full value from our data, we need much more information on that data.

What is possible?

This paper asserts that defining data management solely as a cost-management problem is too narrow a view because doing so artificially constrains any proposed solution's scope, unnecessarily restricting the use and limiting the value of life sciences data.

Metadata

Metadata belongs at the core of any comprehensive data management strategy. To manage data properly, we need detailed information about every piece of data. While basic file metadata was sufficient for legacy filesystems 30 years ago, it does not help navigate collections of multi-petabyte storage repositories. Rather than providing just the path, owner, creation, modification, and last access dates, plus read/write permissions, we need detailed information that includes:

- Detailed description of the data in the file;
- A description of the machine that generated the data, if machine-generated;
- The last calibration date and references, if machine-generated;
- The research and project file associated with the data;
- Any funding or cost center code – which may have associated retention and sharing policies;
- Security protocols required for the file;
- Governance protocols applicable to the file;
- Protection protocols: Backup, number of copies, off-site;
- An expiry date/destroy date.

This kind of information provides **context** – not just about the file itself, but also full details for all the regulatory, security, and other requirements that best-practice data management must achieve.

If a legacy filesystem were to be completely redesigned based on the requirements, challenges and volume of data we are working with today, and there were no requirement for backward compatibility, this is what a re-imagined filesystem could look like (Figure 1). Instead of our primordial and limited filesystem metadata, we'd have a far more comprehensive metadata capability.

Persistent, unified data depositary to eliminate data silos

Metadata would drive all the storage, security and governance protocols for the underlying content data, which can even be distributed across the world in separate storage repositories at multiple sites. In order to accomplish this, the metadata must reside outside of the file itself, existing as its own persistent, unified data repository that is searchable across the entire organization.

Separating the metadata from the data enables rapid, ad-hoc search across the entirety of data. From a local

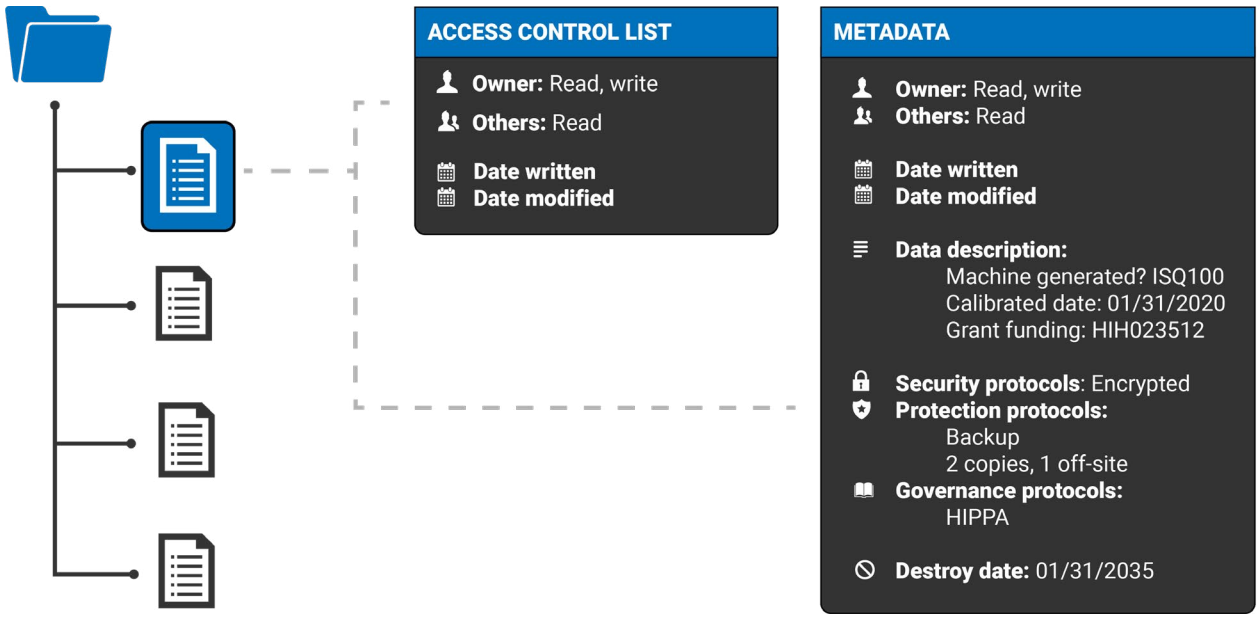


Figure 1: An example of a file system with comprehensive metadata capability.

PC, any authorized researcher can search: all data recently created and currently residing on the fast first tier of storage; the less active data on slower, less expensive storage; the long-term archive in another location, potentially the cloud; and even data residing in deep archives (such as on tape) in a data warehouse.

Applying security or governance protocols also becomes more flexible when metadata is separated from data. For example, if certain data are subject to HIPPA protocols, other researchers may not have access to the files themselves, but could potentially "see" that those files exist, and what kind of data is available on them, without needing to access the files themselves, or view any of the associated metadata that is restricted by protocols. Or, in the case of 'black operations', only a few users would even be able to see any metadata associated with 'black files'. It would be as though these files did not exist for users who lack the appropriate credentials.

This metadata capability would have a tremendous impact on research, where the ability to easily search across all data, regardless of where it exists, is often invaluable. Many researchers currently struggle just to find the data they've recently generated, let alone combing through a mess of files to find related data.

Consider a hypothetical researcher investigating a particular form of cancer. Before developing a series of experiments, the researcher searches all the available metadata – not just data they have created, but all data across the entire organization related to this particular type of cancer cell. This research is on the cutting edge, so it's unlikely they find the exact experimental data they were planning to develop. However, perhaps the researcher finds a significant amount of related data, which they can then review to better refine the new experiment design. Or perhaps after reviewing it, the researcher finds that while the data isn't exactly what they were looking for, it does provide additional information on the cells they were planning to target. It doesn't confirm the new hypothesis, and therefore the researcher can decide to pursue a different and more promising line of research. This ability to search across all existing data in an organization not only reduces time to market by enabling researchers to better contextualize and refine their research, but it can also potentially reduce time spent pursuing less optimal paths and can help to avoid accidental duplication of research.

This isn't as counterintuitive a process as it may seem. Anyone who's visited a library in the past hundred years will have used a very successful analog equivalent – the Dewey Decimal Classification system. Dewey index cards contain metadata about a book; those cards are in a separate store; and often the referenced content – the book itself – can be at one or more of several sites. The index card helps users search for and retrieve it. The Dewey Decimal system has also been successfully digitized – which makes a similar metadata system for an organization's data holdings seem like a prescient choice.

Data tiering

Continuing the library analogy above, imagine a large inner-city library that spans many floors, each with many rooms. If it's well staffed with professional librarians, you might give a librarian a few key search points and they might return in a couple of minutes with your books. If the library is staffed by volunteers, they might tell you on what floor, room, shelves and so on you can find your relevant books. If the library is short on staff and it's all 'self-serve', the user might go to a catalog terminal to search for various terms organized using the Dewey Decimal system, and then walk over to the destination to find the appropriate book.

If space is at a premium in this library but it still aims to provide the best 'quality of service' to its users, it might sensibly keep the most frequently used books in the most accessible place – perhaps close to the front door, or where the search for books is carried out. Perhaps the library strives to keep a good archive, maybe a unique copy of all books and journals they have. Esoteric journals that haven't been read for a couple of years might be best stored off-site in a warehouse where they are still accessible, but it takes some time and effort to get the desired volume; a patron has to walk over to the warehouse, get the ladder, then go to the top shelf. The library would consider this a wise choice: it's much cheaper to store those journals in the warehouse than in the downtown library.

All of the above concepts have parallels in the digital domain. Storage tiers are often characterized by the underlying technology, but what's really important is the 'quality of service' – measured with attributes such as access speeds, cost, reliability, and so on. Similarly, data availability may be 'self-serve' or require 'help' – which could be a software application or tool in the digital domain.
Metadata significantly enhances the effectiveness of data tiering and archiving strategies, compared to data-tiering based strictly upon basic filesystem attributes. Currently, most data tiering is based solely upon the timestamps within legacy file attributes. If a file hasn't been touched in x-hours, it tiers to a slower disk. If it hasn't been accessed in y-months, it moves to archive. Once archived, that data would typically no longer appear in file searches.

However, it's easy to start managing the life cycle for the data file by policy. We only need a little more information about the file, such as:

- The project the data is a part of;
- Governance requirements; and
- Retention requirements, and other pertinent information,

With full confidence that the owner has annotated the correct disposition in the metadata, the file's content is easily migrated to the lowest tier of active media, or even comfortably pushed to deep archive or tape, or even data destruction. Regardless of where the file contents are located, the file would still appear in metadata searches and be useful to the organization. Unfortunately, traditional data management is far more ad-hoc – "we're out of space on this drive, so let's move a bunch of this older stuff over here and deal with it later" – but all too often, "later" never comes.

Automated data workflow and curation requirements

The biggest objection to data management through metadata is really an objection to the labour of creating the metadata in the first place. Data managers argue that end-users will simply refuse to curate their files because it takes too much time and effort. This would be a valid objection if each user had to individually apply the description, security, governance, and project codes to each file – which is certainly too onerous a task to demand! However, data management platforms can automate the majority of this work.

Using metadata 'scraping', a workflow could be as simple as creating a project directory and ensuring each file and subfolder inherits the project's metadata tags. Then, the user would only need to add metadata specific to that file. In the case of machine-generated data, the machine itself might provide much of that metadata, in combination with folder-inherited metadata fragments. Automated workflow could transform what used to be the most obscure, opaque data in the organization into the most transparently documented.

Let's explore how this might work. A researcher is setting up a sequence of samples to be imaged on a device that they are timesharing with other users. The researcher pulls up the scheduling app and enters the sample number sequence. Their personal information is tied to authentication, and then they enter specifics about the project, cost center and sample data. This data becomes "user-defined" metadata. When the samples are imaged, the system automatically applies metadata including the machine model number, calibration table and various environmental settings. This data is extracted automatically and paired with file information including creation date, size, type, et cetera, which becomes the "system-defined" metadata. The imaging data is collected from the lab machine

and transferred to the next stage of an analysis pipeline, and the resulting files are tagged with that same metadata. Already the researcher's data is easier to find, and it's easier to replicate the experiment. If the imaging data was collected on a machine that was suspect or out of calibration it can quickly be discarded, too.

Other workflows often have ancillary collaborating systems such as a LIMS (Laboratory Information Management System). A helper application could add data from the LIMS to the metadata for easier discovery, search and retrieval at a later date.

Data provenance, security and governance

Filesystem attributes allow for only very coarse classification of data – who owns what, who has permission to read/write, and its date of creation. Because there's so little information about the content data, it seems prudent to restrict all data to protect certain governance requirements on some small subset of the data. For example, if some data in a laboratory environment is subject to HIPPA restrictions, and specific files require limited information visibility, we might decide "better safe than sorry", and treat all information in the lab as HIPPA data. While that may reduce the risk of failing to comply with HIPPA requirements, it could also adversely impact research by needlessly restricting data access to non-governed data.

Data governance

The Data Governance Institute defines data governance as "a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models, which describe who can take what actions with what information, and when, under what circumstances, using what methods." Research institutes are typically subject to certain data governance requirements, including HIPPA, GDPR, or other state- or federally mandated privacy laws. It's a serious matter to determine who can see which data, and to prove that data has been removed as required.

It's also very hard to comply with changing rules and regulations without a flexible data management regime that uses automated tools, applications and services. For example, if a hospital moves all patient information from July to tape storage, but a later change to privacy regulation requires this information to be permanently erased, how does the hospital prove that the appropriate records have been recalled and erased from its archive?

Data provenance

Data provenance is a critical feature of robust research. A clear chain of custody of the data demonstrates the veracity of the results and conclusions, as well as enabling other researchers to use the same data and methods to re-create identical results. The National Network of Libraries of Medicine (NNLM) defines data provenance as "the origins, custody, and ownership of research data. Because datasets are used and reformulated or reworked to create new data, provenance is important to trace newly designed or re-purposed data back to their original datasets. The concept of provenance guarantees that data creators are held accountable for their work, and provide a chain of information where data can be tracked as researchers use other researchers' data and adapt it for their own purposes."

A file versioning system is essential to deliver high-fidelity data provenance. The initial file must be locked, and each change to the data should generate another version copy of the data, together with an audit of that file that discloses all versions of that piece of data. This enables researchers to identify who changed the result from, for example, "42.7% of patients" to "about 50% of patients", and when.

Data provenance can be extended even further beyond versioning, to provide a complete data handling history that precisely records any and all changes in the metadata. An image of a sample could include the brand and software version of the imaging device, its last calibration date, and reference the calibration details. If each process were documented as this file moves through the research workflow, we could see the entire flow of the original file through the research processes to the final report.

Data security

Data security is typically a patchwork of tools and processes – a combination of Authentication, Authorization, and Accounting (AAA), ACLs, and bolt-on data encryption and Data Loss Prevention (DLP) tools. Data encryption tends to be applied at a directory or storage level – either an encrypted device or area is set aside, and users or processes put files in that area. This is a workable solution but it essentially requires users to save files and copies to the appropriate location.

If the process were metadata-driven, each file would self-identify as required by certain security protocols. Rather than trusting a DLP tool to identify an image of a chemical or biological description as being proprietary, and then to block it from leaving the network, the metadata itself would identify that the file requires encryption and apply the appropriate DLP protocols.

Data protection

Not all data is created equally; the more valuable the data, the higher its level of protection. Every organization defines its data protection levels, data protection schedules, data recovery point objectives (RPO), and data recovery time objectives (RTO) – and these protocols could easily be listed as metadata. The protection program could then use this metadata as its template driver – for example, by ensuring that there are two copies on-site with a third off-site, and that it is backed up or copied in a manner consistent with the designated RPO and RTO.

Next generation filesystem

Taken together, the data management processes we have so far described constitute a next-generation filesystem. While legacy filesystems have so far sufficed to ensure the integrity of the bits and bytes associated with a file, they were never designed to manage the incredible volume and velocity of data that many organizations are coping with today. By contrast, a next-generation filesystem could: enable global search across all the data within an organization; provide detailed versioning and audit; permit easy specification of the data's required security, protection and governance policies; enable metadata curation; and facilitate automated workflow.

However, a next-generation filesystem by itself would not lead to data nirvana. For that, we need to add powerful data management processes and practices.

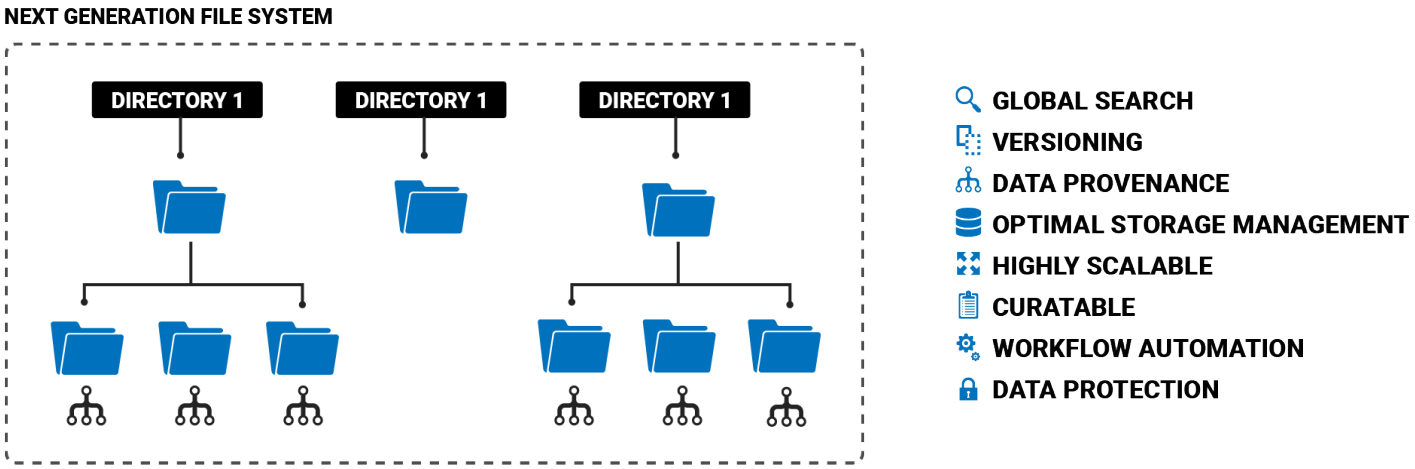
Other aspects of research data management

Sharing data across disparate sites

Today, research is often done collaboratively and across geographically remote sites. A robust data management system needs several key capabilities to perform efficiently as a distributed system.

First, it needs a single "glass door" into all the data a researcher can access. Presenting all data via a "single global namespace" is an emerging approach that can achieve this. Such a global namespace works like a phone directory on a smartphone: if someone wanted to call Mary Jones, they would type "m", "a", "r", and so on into their phone, or tell the phone's digital assistant: "Call Mary Jones" and it happens. If they want to be more specific, they might say, "Call Mary Jones home phone", in which case the phone number is abstracted from the user.

It didn't always work as smoothly as this. People could once remember many phone numbers in their youth; but over time, as people moved or changed their



carrier, those phone numbers that were burned into their long-term memory are now worthless. They were a physical address, in much the same way as knowing that lab126.dat is located on sally:/usr/people/sally/aug-runs/big_data/lab126.dat. Over time, that file likely will be moved, and then even Sally will have a hard time finding it, even though it is Sally's file!

But if Sally could tag the lab126.dat file with useful data – that is, metadata such as "Frank Jones right knee MRI", "Aug 6, 2020" or "Tumor" – she and her colleagues could much more easily find and use this data file in the future. Furthermore, after time passes and this file is moved from primary storage to secondary to archive, anyone can still find it at the same place in the global namespace, even though its physical location may have changed several times.

A single global namespace gives researchers the capability to share data with colleagues across town, between states or even internationally. And as an added benefit, federating data silos under one global namespace usually makes storage significantly cheaper, because the system uses less raw storage, duplicates less data, and saves on egress charges from cloud storage.

HPC processing

Running complex experiments or simulations often requires specialized and dedicated High Performance Computing (HPC) resources, which are typically shared between departments, or even rented from another organization or from the cloud. Typically, HPC environments comprise clusters of massive, very bespoke multi-processor machines, directly coupled with dedicated storage that can perform fast enough to keep up with the host's I/O requirements.

This dedicated storage is called "scratch", and typically data is copied from "persistent" storage to "scratch" storage for processing as a "job" on the HPC device. Because scratch storage is so esoteric and expensive – and often tuned to work well with the accompanying HPC resources – space on it is usually rare. Once the HPC job is executed, the input and output (result) files are copied from the scratch storage to make room for the next batch of HPC jobs. This whole process of getting the appropriate data onto scratch and then back off it again is sometimes affectionately known as "data wrangling".
The right data management platform should automate this process as much as possible, adding appropriate metadata at every step in a data processing pipeline. Such a data-wrangling approach makes it readily possible at some future date to see what inputs produced what results. It also maintains the ability to reproduce the 'run' exactly, as it preserves all the appropriate input and output files and associated parameters.

Conclusions

Data management can be a daunting challenge in today's life sciences research environment, but you are not alone – instead, you are part of a very large community who all face similar challenges in dealing with the volume and velocity of today's research data. This paper summarizes the common problems in data management that we have quickly identified in discussions with our peers. Some of these problems are easy to fix – for instance, by moving data by policy to a cheaper storage medium.

Other problems are more subjective and organization-dependent, like how much time a scientist spends trying to organize or find the appropriate data for their next task. It's harder to quantify the intangible benefits that accrue from enabling a researcher to spend more of their time on research and less time finding things or picking up after themselves because of a disk quota or system limitation. Indeed, many researchers have told us that they want ready access to their data and that of others in the organization, without having to become an IT expert.

The authors would greatly appreciate feedback and suggestions about your research data management issues. Based on feedback we already plan to write further papers on topics such as data curation, automating data management pipelines and data lifecycle management, as well as on challenges such as consistent ontology, and adherence to process, standards or governance restrictions. After many laborious years, we feel able to solve many of the issues we have described above, and we invite readers to learn more on our respective websites.

Arcitecta
talk@arcitecta.com
+1 720 583 4006
www.arcitecta.com

G-Know
info@g-know.com
858.356.5790
www.g-know.com

Arcitecta®, Mediaflux® and XODB® are registered trademarks of Arcitecta IP Pty. Ltd. in the USA and Europe and trademarks of Arcitecta IP Pty. Ltd. in Australia.

All Rights Reserved. © Arcitecta IP Pty. Ltd. and G-Know Inc. 2022-2025.