The ultimate guide to starting your digital preservation journey

An introduction for custodians of valuable digital assets to the basics of long-term archiving and preservation.

arkivum

Bringing archived data to life

Contents

03.
Introduction

04.
What is Digital Archiving & Digital Preservation?

08.
Where to start?

14.
Sustainability

16.
Building the case for digital preservation

18.
Inhouse vs Outsourcing

20.
Choosing a provider

22.
Glossary

Introduction

Higher education institutions (such as universities) and cultural and heritage institutions (such as galleries, libraries, archives and museums (GLAM)) hold some of the world's most valuable and historically important documents and assets. Not only do these need to be protected for the long-term, but they must be accessible and usable for future generations to come.

While many of these organisations have embraced, at least to some degree, digital technologies to support their archiving and preservation requirements, challenges still exist. In **our research**, we found that 62% of institutions face the challenge of having data stored in multiple locations, 40% are unsure of where to begin in relation to adding metadata and 31% face the challenge of digitising their content.

It can be argued that the COVID-19 pandemic has accelerated the requirement for cultural and memory organisations to catch up and become digitally accessible but knowing where to start and understanding the right route for your institution can be difficult. Part of your challenge may be in getting the wider organisation to understand why a digital solution is needed or how to establish the appropriate mix of people, processes and technology to ensure that your data remains secure, searchable and accessible.

We have written this eBook to help equip those working within GLAM and Special Collections and Archives with the knowledge to begin a digital preservation journey or improve on existing processes in place. It will also help you understand how you can overcome certain challenges and even, what and when you should archive.

What is digital archiving and preservation?

Digital archiving and preservation can often be confused with storage backups so it's important to understand the differences between them and what their purposes and outcomes are.

What's the difference?

Fundamentally, a data backup is where a snapshot of your data is taken and stored elsewhere so that if required, you can recover this data from a certain point in time. Typically, this recovery method is used in disaster situations, such as an incident of accidental deletion of files, hardware issues or server failure.

Typically, daily backups are kept for a certain number of days...which are then replaced by weekly backups... which are then replaced by monthly backups. Leaving your data in a live system long-term with backups running on a continual basis can lead to a lot of data being stored, opening you up to risks of data loss, excessive costs and/or data deletion. Relying on a backup for a long period of time does not protect against corruption or from formats becoming obsolete.

That's not to say that backups aren't important, they are. They are simply not designed to ensure the long-term integrity and preservation of your data whilst also providing access to current and future students, stakeholders, and researchers.

By comparison, a digital archive is an ongoing, managed environment that focuses on the preservation of your data - forever. It goes beyond creating a copy and instead, focuses on the accessibility and re-usability of your data.

For example, a student many years into the future undertaking a research project or thesis, could search and access materials held by your institution – whether that be research data, digitised copies of transcripts or research papers into the evolution of humanity – and use this to form part of their paper or project. Without having a database of these historical digital assets in the first place, future students and researchers may not be aware of artefacts and data in existence.

The benefits of digital archiving and preservation

- **Guarantee 100% data integrity:** Safeguard digital assets and provide peace of mind that they are safe and secure from becoming lost, corrupt or hacked.
- **Removal of data silos:** Consolidate your resources.
- **Long-term accessibility:** Ensure the right people can find and have access to the right data when they need it.
- **Usability:** Digital preservation protects against file obsolescence.

The three main elements of digital archiving and preservation are:

1. Searchability

Ensuring that the information you need can be easily found in the future. A key feature in enabling this is by assigning the appropriate metadata against each asset. There are two types of metadata worth noting:

- **Descriptive metadata** (title, author, date of publication, description etc.)
- **Technical metadata** which describes technical properties of a digital file or the particular hardware and software (where the data resides, the structure of the data etc.)

Let's take a photograph as an example. An institution can store a digital copy of a photo and have accompanying details to search by, including when it was taken, who by, where it was captured and perhaps, details regarding what it is of.

With potentially limited knowledge, they'll be able to search an abundance of files which fall within these search criteria to be met with relevant results of matching photos...all because of metadata. Years into the future, someone can easily find that photo for a number of different reasons.

2. Accessibility

Once your asset or data has been located, accessibility ensures that it can be opened and read. Effective digital preservation ensures that valuable information is stored in multiple locations, regularly checked for validity and easily accessed by those who need it (and with the correct permissions), when they need it. Essentially, this means that your artefacts and data will be kept alive and shared with the world where appropriate to continue educating and informing.

For example, a university may upload its research data into a campus-wide repository so that all stakeholders can access it, yet there are a certain number of projects which must only be accessed by those who are directly involved. A digital archive can ensure that those particular stakeholders have the correct user permission to the associated files, and all other general stakeholders (who have access to the campus repository) won't be able to access it.

3. Usability

Formats and technologies are in a constant state of evolution. Equipping your institution with automated digital preservation tools (such as file normalisation) ensures that your data is regularly checked and updated to the latest and most appropriate format so that those with permitted access (if applicable) can continue to make use of your data.

Below are two real examples for how a digital storage solution has helped world-renowned educational institutions:

Princeton utilised a digital archive to manage their growing amount of born-digital content across a variety of file formats and policies. These challenges were exacerbated by strict security requirements for handling personal data (such as student records) and complex user access policies dictated by the university's information governance policies.

The Irish Traditional Music Association (ITMA) needed to ensure accessibility to its valuable sound recordings, still and moving images, manuscripts, printed materials and other irreplaceable assets, as well as safeguard and preserve them for the long term.

"After a number of years refining scalable workflows for the ingest of assets into Arkivum, ITMA is now well ensconced on its digital preservation journey. We are safeguarding an extensive digital collection while also providing meaningful access to a national public archive for Irish traditional music, song and dance," **Maeve Gebruers, Archivist Irish Traditional Music Archive.**

Where to start?

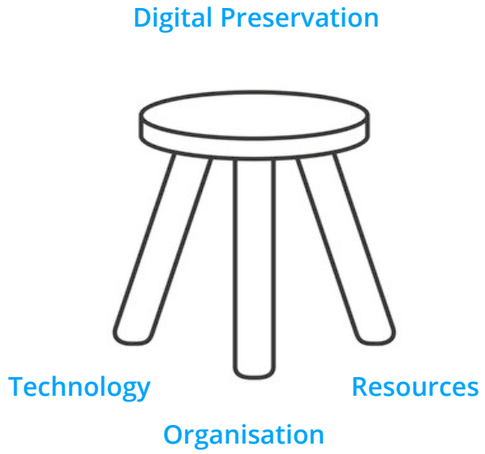
Not every organisation needs an incredibly complex digital archiving and preservation approach. Instead, every organisation must find the appropriate approach which fits their requirements.

There are various models you could choose from to help get you started but for now we'll look at two: the three-legged stool model and maturity models.

Three-legged stool model:

To achieve a successful and sustainable approach to digital preservation, technology, organisation and resources must be considered in tandem. This model demonstrates that without the consideration and maintenance of all three elements, a digital preservation programme will ultimately collapse. It's therefore a balancing act to ensure that the correct provision and attention is directed towards each element – or leg of the stool.

1. Technology
2. Organisation
3. Resources



The technology leg represents the necessary hardware, software and secure environments required to sustain a digital preservation programme. Some areas to consider within this element include:

- Storage and back-up
- Repository and preservation systems
- Security measures

The organisation leg of the stool looks at the elements required to address the organisational needs and can help drive the organisational change required for a successful digital preservation programme. Some areas to consider within this element include:

- Policy and strategy
- Workflows
- Risk registers & benefits models

The resources leg looks at the time, money and people requirements of a digital preservation programme. These are the resources required to create and maintain a sustainable programme and can include:

- Business planning
- Funding
- Staff skills
- An organisational commitment to sustainability.

Maturity models

Whether you are starting out on your archiving and preservation journey or are unsure of the effectiveness of your current approach, it may be helpful to look at assessment frameworks like maturity models. These tools are designed to assess the effectiveness of your current approach and/or processes.

Generally, maturity models consist of several levels and the higher you go, the more mature you are. These tools are useful in understanding your existing level of maturity (where you are now) against where you want to be and formulating a plan in achieving this level increase.

They equip you with the knowledge to understand:

- The practical capacity of your organisation.
- Understanding your organisation's goals and missions.
- Knowing a little about the digital assets and data that require archiving.

Forget a one size fits all approach. These tools help you understand where you are as an organisation and map out where you want to be.

Which maturity model?

There are various models you could look at for inspiration and guidance but the below may be a good starting point.

The National Digital Stewardship Alliance levels of digital preservation (NDSA levels) (more commonly used in the US) is a simple one-page guide with four levels of good practice that covers functional areas of archiving such as storage, integrity, control, metadata and content.

The Digital Preservation Capability and Maturity Model (DPCMM) provides 5 levels of maturity for 13 capability areas that cover all aspects of digital preservation infrastructure and services.

The Digital Preservation Coalition Rapid Assessment Model (DPC RAM) (more commonly used in Europe and Australia) is a new resource from the DPC and builds upon other maturity models including the NDSA levels and DPCMM. DPC RAM is notable by being designed to be simple to use and includes a self-assessment worksheet and graphical analysis.

Leveraging data management best practice

The data generated by, and the collections and archives held by GLAM institutions can often aid and advance research and cross-institute collaboration. However, in order for these assets to be utilised, an institute can follow (whether loosely or strictly) data management guidelines with the most widely adopted being those of the FAIR practical principles.

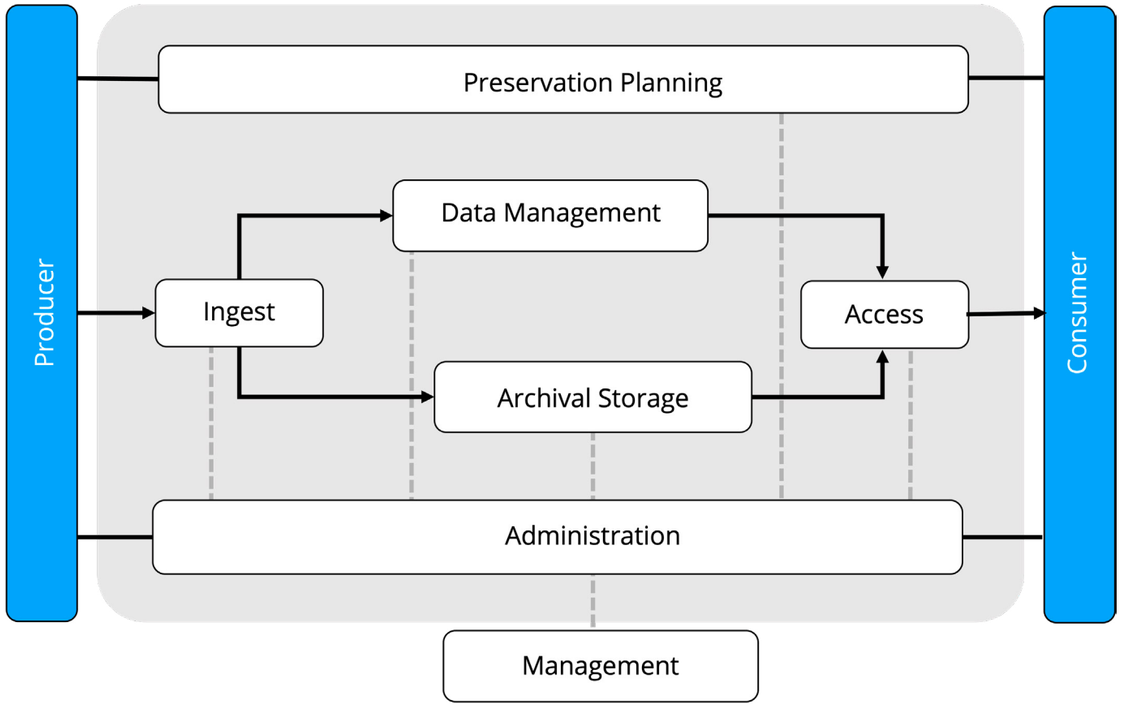
These are a good basis to align your data management with, and comprise of four elements to help identify if you are storing your data in such a manner that it remains:

- **Findable:** The first step in (re)using data is to find it. Metadata is essential for automatic discovery of datasets and digital assets.
- **Accessible:** Once the item or dataset has been found, the user must be able to access it.
- **Interoperable:** The data usually needs to be integrated with other data. In addition, it will need to interoperate with applications or workflows for analysis, storage, and processing.
- **Reusable:** The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

These guidelines have been designed to act as a tool to help an organisation better manage their data, improve their digital archiving and preservation efforts and most of all, help ensure data longevity.

The OAIS reference model

The Open Archival Information System (generally known as the OAIS reference model) has become the go-to framework for digital preservation. An OAIS is an archive consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available to a designated community. The standard defines a set of responsibilities that an OAIS archive must fulfil, and this allows an OAIS archive to be distinguished from other uses of the term archive.



Sustainability

We're generating more data than ever before and it's therefore no surprise that sustainability within digital preservation is a growing concern.

In 2020 alone, each individual produced 1.7 MB of data every second. Not every day, but every second and there are 86,400 seconds in a single day...

It has also been estimated that 90% of data was created in the last two years.

There's no reason to feel overwhelmed or disheartened at the prospect of how your archived data could be impacting the environment. Organisations need to store data, and some produce or preserve far greater quantities than others. What matters is that the data stored is done so in a sustainable and efficient manner.

...simply altering how we use technology is not the solution. Rather, we must critically examine through the lens of sustainability how much we preserve, how we preserve it, and how we make it available to our communities.

Before you begin your archival and preservation journey, it's vital that you consider the below questions early on to build a foundation for your plan:

- Do you need to store this data?
- How accessible does it need to be?
- Does it need to be easily retrievable?
- What quality must it be archived in?
- What does the accessible copy need to look like?

Everything you archive and preserve has a cost associated with it, whether that be the physical offsite overheads, human resources or the energy used. There's no need to archive every document you possess, or every draft version of a piece of work if you don't need to.

Accessibility determines the level of storage you require. For example, video and photos should generally have original larger files in deep storage with lower quality, access versions being readily available. There's no need for you to make original video and image files easily accessible when a far lower quality version will suffice for online distribution.

On the flip side, a deeper level of storage could be used for raw data as it's unlikely you'll need quick access to it. It may be the case that you require easy retrieval for an access copy which is a lower quality than the original but is cheaper to archive, has greener credentials and quicker recovery times.

Another fact to consider is, when looking at potential vendors, how environmentally friendly are they? For example, Microsoft have set themselves the goal of becoming carbon negative by 2030 and it's highly likely many other public cloud providers will look to achieve the same scale of green credentials.

Taking the time to address the above questions and undertake a maturity assessment will help you devise an appropriate storage plan for your institution's data.

> *"... environmental sustainability of data-centric activities, including digital preservation, is undeniably something we should all be concerned about,"*
>
> Matthew Addis, CTO, Arkivum.

Building the case for digital preservation within your organisation

Digital preservation entails far more than simply storing your data and it may be that you need to present the case of the additional value a digital archiving and preservation solution can provide your organisation with.

In a recent survey, we found that 60% struggle to obtain buy-in from their wider organisation and 64.4% struggle to obtain an appropriate level of investment.

We know that many archivists and records managers are concerned with the appropriate long-term management of their data, but they do not necessarily have the support of the organisation behind them to achieve best practice. Perhaps the biggest one is you're investing in something which is primarily seen as preventative, and the value of this investment will only be realised if something goes wrong. As such, it can be difficult to prove the ROI of this type of investment in comparison to other more tangible budgetary spends which may be more easily measurable, such as an increase in efficiency or time management within a department.

Therefore, how can you demonstrate and communicate the requirement for a digital archive and preservation solution?

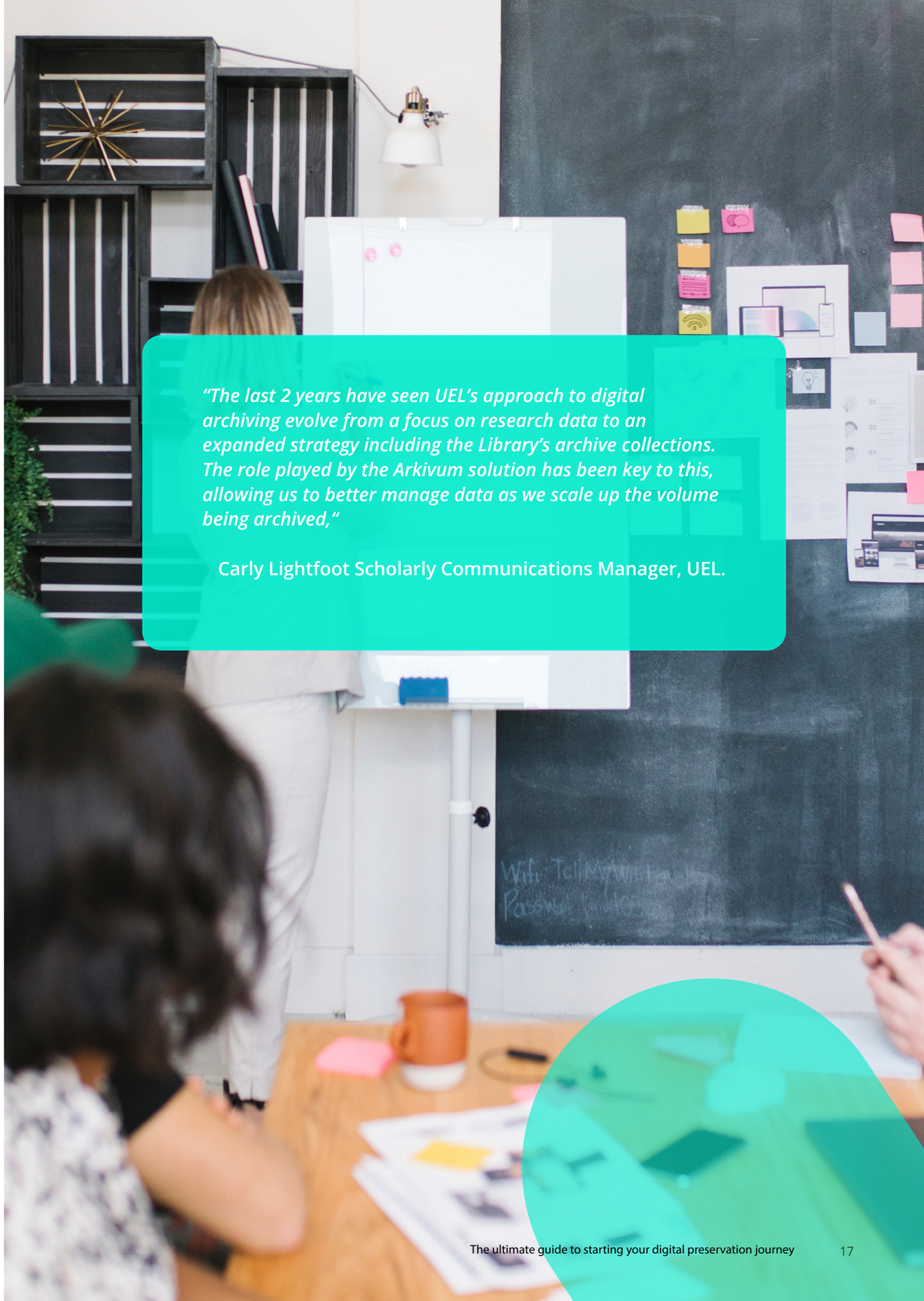
Every organisation is different, has different objectives and faces different challenges. To build your case, it's worth considering how you present digital preservation and the additional value your organisation could benefit from.

As mentioned, the greatest challenge to defining the value of digital preservation is that the true value is often not realised until something goes wrong when proper investment has not been made, and information is lost.

It's important to shift the conversation with the wider organisation and present the process of digital preservation in a different manner. Instead of viewing it solely as a defensive mechanism to maintain old records, it is something to deliver strategic value.

One example that could add value is the maintainenance and leveraging of corporate memory, through the effective preservation of previous processes, activities or campaigns. As people stay in jobs for shorter periods than ever before, there is a growing risk of duplication of work or missed opportunities for the business.

Organisations could potentially leverage this knowledge for competitive advantage by ensuring valuable knowledge is passed on and value assets can be reused or repurposed.



"The last 2 years have seen UEL's approach to digital archiving evolve from a focus on research data to an expanded strategy including the Library's archive collections. The role played by the Arkivum solution has been key to this, allowing us to better manage data as we scale up the volume being archived,"

Carly Lightfoot Scholarly Communications Manager, UEL.

Inhouse vs Outsourcing:

Whether you're starting out, or are considering moving from an in-house operation to an external solution (or possibly vice versa), we've outlined a few things worth knowing to help you make the best choice.

Opting to keep the act of archiving and preservation in-house is sometimes preferable because it can feel like a safer option – if you can see it (virtually, of course) then you can protect it. Depending on how much data requires archiving, it can sometimes be a cheaper alternative to using a third party. Additionally, an in-house route would provide you with complete control and provides the option to customise the build to your requirements.

However, there are a few challenges and concerns to consider. An in-house solution is reliant on extending the responsibilities of personnel and relying on their capabilities. This could be a cause for concern if they have limited or non-existent technical ability and experience in digital archiving, as well as placing additional pressure on their role. This option can also be expensive to maintain and even more so if changes or upgrades are required.

Using an external vendor can alleviate these issues and deliver an efficient medium to store your long-term data. It's also worth noting that if you choose to use a third party, your data is still your data. Opting for an external solution does not negate the control and ownership you have but you should ensure that your data is easily retrievable from their systems is required (e.g. Escrow).

Choosing a provider

Before you outsource your solution, it's imperative that you choose the right provider but making this decision isn't always easy or simple.

Outlined below is a checklist of questions and talking points we'd suggest using when speaking to a third party:

- ☐ What protections do you have in place to stop data being tampered with?
- ☐ How do you provide access to specific users?
- ☐ What processes do you have in place to mitigate against data loss or corruption?
- ☐ Do you have anything in place to preserve data? (I.e. maintain preservation copies of each file)
- ☐ What expertise does your team have in digital archiving and preservation?
- ☐ Is the solution aligned to best practice approaches such as FAIR and/or ALCOA+?
- ☐ Are there any additional costs associated with accessing my data?
- ☐ If I need to, how easy is it to take all my data from your system?
- ☐ How sustainable is your solution?
- ☐ What systems do you integrate with?
- ☐ What approach do you take towards security and quality? For example, are you ISO 9001 and ISO 27001 certified?
- ☐ How can you meet industry regulations? (If applicable).

From asking these vital questions, you will be equipped with knowledge which will help you to better understand their solution and business, and how this will align with your company's data archiving requirements.

Glossary:

Authentication: A mechanism which attempts to establish the authenticity of digital materials at a particular point in time. For example, digital signatures.

Born-digital: Materials that originate in a digital form.

Digitisation: The process of creating digital files by scanning or converting analogue materials. The resulting digital copy is then be classed as digital material.

Escrow: A legal concept whereby an asset is held by a third party on behalf of two other parties.

Fixity check: The process of verifying that a digital object has not been altered or corrupted.

Metadata: The descriptive text associated and indexed against uploaded assets and data to make them searchable.

OAIS: Open Archival Information System - any organisation or system charged with the task of preserving information over the long term and making it accessible to a specified class of users.

SaaS: Software as a Service – a software delivery model in which software and its associated data are hosted centrally (in the Cloud) and are accessed by licensed users using a web browser over the internet.

Three-Legged Stool: A conceptual approach to digital preservation that suggests a fully implemented and viable preservation programme addresses organisational issues, technological concerns, and funding questions, balancing them like a three-legged stool.

About Arkivum

Arkivum is recognised internationally for its expertise in the archiving and digital preservation of valuable data and digitised assets in large volumes and multiple formats. The long-term security, integrity and accessibility of data is crucial for all Arkivum's clients and partners, who share a commitment to good practice in its stewardship and governance.

Arkivum's specialist software and services are chosen by major institutions and commercial organisations in a diversity of sectors, including life sciences, research (CERN), financial services, and organisations in higher education, culture and heritage. Confident in Arkivum's reputation and resources, they are in a position to maximise insight and discovery by deriving optimum long-term value from their data, collections and intellectual property.

Headquartered in the UK, with presence in the US, Arkivum advocates the use of the FAIR principles in data management: Findable, Accessible, Interoperable, Reusable. Arkivum is also certified in ISO 9001 and 27001.

For more information about the product, or to arrange a demo, please email us on hello@arkivum.com or visit www.arkivum.com

@Arkivum

Address: 450 Brook Drive, Green Park, Reading, RG2 6UU

arkivum
Bringing archived data to life