Tessella Safety Deposit Box

The Case for Digital Archiving in Life Sciences

By: John Russell, Contributing Editor, *Bio•IT World*

Produced by Cambridge Healthtech
Media Group Custom Publishing

Tessella
Technology & Consulting

Tessella Safety Deposit Box

The Case for Digital Archiving in Life Sciences

Even as the Era of "Big Data" engulfs life sciences and researchers rush to wring immediate value from the electronic data deluge, the long-term archiving of this valuable asset is unfortunately, widely neglected. Few industries face a more daunting data management challenge. The size of datasets is numbing, the diversity of data types is bewildering, and technologies used to produce and analyze the data are advancing so rapidly that coping with obsolescence is becoming a more urgent challenge.

It used to be that only large 'memory institutions' – a national archives, for example – worried about and implemented comprehensive long term archival systems. Now, the surging data deluge (annual worldwide sequencing capacity today is roughly 13 quadrillion DNA bases.) and a growing awareness of the substantial benefits and risk avoidance features that a comprehensive digital archiving solution can deliver, are prompting many industries to re-think their long-term data archival strategies.

Within the biopharmaceutical sector, tackling the archiving challenge is complicated by a legacy of decentralized data and content repositories, often at the research group or department level. These dispersed data repositories may meet immediate research and operational needs but may not be aligned with corporate goals or in a form that enables long-term data integrity, authenticity and accessibility. However, the return on the investment and effort to implement centralized long-term archives can be substantial. Consider just a few of the business drivers prompting life science organizations to adopt comprehensive archiving solutions:

- **Cost reduction.** Multiple research information systems are often kept running when they are no longer actively used just to house and maintain access to the data. Adoption of a centralized archive platform allows these systems to be decommissioned, eliminating licensing and operational costs.

- **IP protection.** IP and supporting data (experimental & analysis) are the lifeblood of the biopharmaceutical industry. Proper archiving safeguards the data (authenticity, integrity and accessibility) and is critical for defending patents and informing collaboration and licensing deals.

- **Data repurposing.** Given the speed with which tools for drug discovery & development are evolving, there is an enormous opportunity to have legacy data readily accessible for re-examination with new analysis and data-mining technologies to support new drug discovery efforts.

- **Regulatory compliance.** Perhaps self-evident, complying with the maze of regulations around data handling and retention is difficult and penalties for non-compliance can be severe. For example, FDA requires holding clinical trial data for many years beyond the lifetime of the approved product. An archive platform can ease management of this task, keeping authentic records safe, secure, and readily available throughout the entire retention period.

- **Litigation support.** It goes without saying that biomedical research and healthcare are subject to frequent eDiscovery requests and subsequent litigation. (e.g. Vioxx). These actions can literally wreck a company. Archival systems keep authoritative and authentic data safe and make it much easier to locate and gather necessary information than when the data is dispersed. It's not unusual for a major pharmaceutical company to have hundreds and sometimes thousands of open litigations at the same time.

WHAT IS A DIGITAL ARCHIVE?

It is perhaps useful to broadly define a long-term digital archive and distinguish it from other storage options. A common misconception is that backup and recovery constitute an archival solution. They do not.

A digital archiving solution must provide mechanisms to not only safeguard the integrity and authenticity of content, but to also ensure continued access throughout the entire retention period. The key concept in long-term data archival is this: when content and data are still subject to usage, change or versioning in the course of conducting ongoing business, they are considered 'operational data' and are not necessarily archived. Data and content that won't change and has potential use beyond its 'operational' life – ten years is a typical threshold – are candidates for long term archiving. Determining specifically which data to archive and for how long are policy decisions made by the archiving organization.

A well-architected archive acts as a central repository maintaining secure access to the data, safeguarding data integrity, demonstrating authenticity and provenance. It should also provide mechanisms for enabling continued access so as the technologies used to create and store the content (applications, formats, storage media, etc.) become obsolete, measures are taken to ensure users still have access to the content. Indeed, there is a well-established ISO standard (Open Archival Information System (OAIS) reference model that defines what needs to be in place to support long-term digital archiving.

There is also a process moving towards an ISO standard that will enable organizations to seek an audit and certification of their archives as trustworthy digital repositories – this could be extremely useful in life sciences when disputes over data provenance and validity arise. Significantly, certification criteria are not just about the technology but also about the security, policy, the governance structure and business processes that an organization must have in place to achieve certification.

HORROR STORIES ABOUND

As researchers have raced to keep pace with the data flood and advancing technology, the proliferation of ad hoc systems has become the rule rather than the exception, sometimes wreaking havoc on efforts to maintain access to legacy data. "The basic issue is there is rarely any kind of enterprise wide long-term archiving strategy," says Mark Evans, Archiving Practice Manager of Tessella, recalling an engagement with one pharmaceutical company.

In this case, chemists were in charge of managing their data. Decades ago they'd used VMS and Unix machines to collect data, off-loading it onto tape. When new technology became available they migrated to it. "So tapes gave way to CDs, which were replaced by DVDs, which were replaced by network shares. When the technology changed, they did a poor job of moving the legacy data forward. There was an extensive amount of data orphaned on legacy media and devices," says Evans.

Finding their data became a huge issue. "The chemist responsible for returning requested data would take weeks to find rather recent data, if she found it at all. Older data was out of the question. She had no idea which CD, network share, etc it was on. The amount of data on media that no one even knew existed was alarming."

No monolithic piece of software could magically solve the problem. It was necessary to write dozens of specialized scripts, physically tear apart over a dozen lab computers, bypass passwords, pick apart the source code of legacy systems and migrate many legacy metadata models. Now the company is broadening its effort to centralize the management of long-term assets.

This client's experience is hardly unique. Such 'horror stories' abound. The pharma-ceutical

industry as a whole, widely criticized for poor R&D productivity, has rushed to adopt new technologies to invigorate thin drug pipelines and to streamline operations. The result has been a surge in data generated and it's been stored in varying formats, using a diverse set of tools and applications with little provision for long-term access or preservation.

DIGITAL ARCHIVES CAN RESTORE ORDER

A carefully drawn archive strategy delivered by an effective archiving platform can substantially mitigate long-term obsolescence problems; however accomplishing this goal isn't trivial, particularly given the paradigm change that switching from paper records to e-records represents. On the plus side, it is possible to do far more with digital data – for example, examine molecular structures in 3D and manipulate those structure to uncover promising (or threatening) pharmacologic properties. But unlike paper records, e-records have an enormous range of technology dependencies (hardware environment, application software, data format, metadata, etc.), which must be accounted for.

Moreover, e-records are relatively easy to manipulate which raises a host of issues. Maintaining data integrity and confidentiality are obvious ones. A less obvious risk, for example, includes the ease with which e-records may be found during a litigation discovery action and therefore subject to disclosure. Indeed, timely disposal of records is often an important consideration and usually part of any individual record's retention schedule. When litigation requests do arise, it's also much easier to put a 'hold' on relevant scheduled disposals using the features of a digital archive.

While the move to deploy comprehensive digital archives in life sciences is relatively recent, Tessella has been deeply involved at the leading age of digital archiving for more than a decade. Its clients include several national archives, such as the UK National Archives, National Archives of the Netherlands, Swiss Federal Archives, as well as private and not for profit organizations including, for example, the Welcome Trust Library, and FamilySearch. Tessella has also been deeply involved in multiple collaborative research efforts to establish practical, credible standards and tools around digital archiving.

A comprehensive review of issues surrounding deployment of digital archives is beyond the scope of this brief paper – in part because the needs of individual archives differ so widely – but Tessella has encapsulated many of the core concepts in a white paper, Long-term preservation of digital content, and also on a microsite dedicated to long-term archiving issues.

There is no "one size fits all" approach, notes Evans, who has led many of Tessella's digital archiving engagements in North America. Although technology choices are important, Evans emphasizes they should be subordinate to clear business goals. Many of the insights gained from engagements have shaped development of Tessella's Safety Deposit Box (SDB), a comprehensive digital archiving platform.

Currently, practices within the pharmaceutical world vary widely and tend to be fragmented. For example, archiving electronic lab notebooks is often problematic. Chemistry-centric users work differently than those in biology disciplines. Different tools and hence file formats can generated. Workflows vary. "I've seen cases where one group was seeking to use digital signatures – an individual attesting to the validity of an entry – while another group was were happy to just print them out and sign them." Far-flung locations, region-specific regulatory mandates, differences within disciplines, and proprietary applications and file formats are all forces which complicate the archiving function within life sciences.

STARTING DOWN THE DIGITAL ARCHIVING PATH

Not surprisingly, deciding what to archive is a critical early step. "My first piece of advice to clients is to define a clear policy that specifies what's going to go into the archive," says Evans. Most pharmaceutical organizations already have retention schedules for various content types, and these retention schedules have usually been informed by particular business goals, future research needs, and regulatory requirements.

"One client we worked with decided anything that has a retention period of seven years or more is candidate for long-term archiving. Less than seven years, they decided they could live with maintaining the operational system in which the data was housed and used." Ten years is also a common threshold. Limiting the number of content format types the archive will accept is another popular practice, although doing so can draw pushback from researchers using proprietary equipment and software – a very common occurrence in life science.

It's also important to determine if you wish to retire any existing systems that are being maintained just to provide access to content. If the answer is yes, examine what the return on investment would be for closing down those systems and moving the content into a centrally managed archive. This can be an important business driver. "One client identified roughly 20 systems that they were just keeping the lights on. They calculated that shutting them down (personnel and licenses) would save enough to pay for development of the archive," says Evans.

Assessing content risk is another important early exercise and helpful in deciding what content to archive. Risk can be based on many factors. Is the content stored in highly proprietary format or is it in a widely used format? Are tools available to render the content? How strong is the vendor support, are the specifications published, and is there a credible roadmap for backwards compatibility?

SDB LOWERS THE BARRIER TO ENTRY FOR ARCHIVING

At some point, of course, the content defined by policies must be entered into a long-term archive and managed. The Tessella Safety Deposit Box archiving platform follows the OAIS archiving reference model whose core elements include:

- **Ingest.** Steps required to transfer items from their current location into the archive in a managed manner.
- **Archival Storage.** The storage of the bulk data (usually files) based on standard storage management tools.
- **Data Management.** Tools to manage the storage of the archive, including metadata.
- **Administration**. Tools to administer the system and access it.
- **Access.** Tool to search, browse and download the content of the archive.
- **Preservation Planning.** The module that manages information so that it can be accessed long into the future.

The one standout feature of SDB is an implementation of the OAIS Preservation Planning module. Although other content and records management systems provide many of the OAIS functions, they all fall short when it come to consideration of the long term preservation requirement.

To maintain flexibility and scalability, SDB employs a service-oriented architecture (SOA). "Essentially, SDB consists of small services that perform simple functions. They are highly cohesive and loosely coupled so they don't depend on each other. What we then do is orchestrate those services together to form processes or workflows and we match those to business processes and business workflows. It's not a one size fits all," says Evans.

SDB was deliberately designed to accommodate organizations of various sizes. "We are trying to

lower the barrier to entry for archiving and develop offerings which smaller institutions can pick up and start using quickly," says Evans. "We always recommend they build out incrementally and don't go immediately for a huge holistic solution.

Once users understand the basics, we then encourage them to think about how they want SDB to fit in with their own business processes. It's also a good idea to engage with a solid number of internal early adopters and learn from that experience because this still is an emerging discipline. There are not many best practices yet."

Consider the many issues surrounding ingest, for example. Once you've appraised your content, it's necessary to think about what's needed to get the content into a form that can be ingested into the archive. That might require connectors to other repositories or instrumentation. It might require functions or services to manipulate that content before it ingested. SDB includes a set of extensible frameworks and API to support these tasks.

Security and validity and integrity are always major concerns. SDB performs virus scanning, integrity checking, file format identification and validation, and authentication during ingest in accordance with policies and business processes defined by the user organization. Establishing a clear policy is critical. For example, digital signatures are increasingly popular in life sciences, "but they are highly dependent on a technology stack and can be problematic when it comes to long-term archiving," says Evans.

FUTURE-PROOFING CONTENT

The ability to future-proof content is a critical requirement for long-term archiving and as mentioned previously is a key feature distinguishing SDB from conventional content management platforms. "If we want to have a chance of accessing content created today, decades into the future then we have to do more than just preserve the bits" says Evans. SDB provides robust long-term preservation capabilities including the ability to generate preservation plans – where you set out what you want to achieve and identify the content you want to act on. SDB can do this 'out of the box.' Here are the three common approaches to executing long-term preservation policies:

- **Migration.** This is moving the content from one particular bit stream coding or file format that you may perceive to be at risk of becoming obsolete to a more current form or more persistent, long term form.
- **Normalization.** This involves moving content of the same type to a standard form. For example, one might have lots of images that are in various image formats and you want to normalize those to one particular format. Normalization is typically performed on ingest, while migration can be performed at any point.
- **Emulation.** That's where you emulate the environment (software, hardware, etc) in which the content was created or used. This can be very useful in life sciences where so many proprietary tools and formats are used.

SDB currently has robust normalization and migration capabilities out of the box. Also, a proof of concept plug-in for SDB has been developed to launch emulation environments and then execute the content within the context of that emulation environment. One issue with emulation is that it does not scale well – think of the hundreds of different environments you might have to emulate. Never the less emulation is gaining popularity for selective use.

It bears repeating that although long-term archiving is a non-trivial exercise, it is also delivers tremendous benefits (knowledge re-use, proof of regulatory compliance, IP protection, etc.). SDB has very many features to ease the process of defining and implementing a comprehensive digital archive. Most projects require some customization and clients may undertake those themselves or use

Tessella or another third-party. Tessella exposes all the APIs and integrations points.

Moreover, there is an active SDB user community. "What tends to happen is users share ideas and experiences. Also, if we develop a specific feature for one user, we generally roll it into the next general SDB release," Evans says. Ongoing development of connectors to widely used content management systems – Documentum and Sharepoint, for example – are priorities.

Technology issues aside Tessella recognizes cost containment is an imperative in the pharmaceutical and healthcare industries today. Work is ongoing to create cloud-based SDB software-as-a-service offering. SDB also offers what's called a multi-tenant approach. "It's a single deployment but multiple organizations or divisions can share the use of that installation. To them it looks like their own and they don't see each other's content. They can have their own workflows tied to their specific business processes, see their own policies," says Evans.

LIFE SCIENCES RACE TO CATCH UP

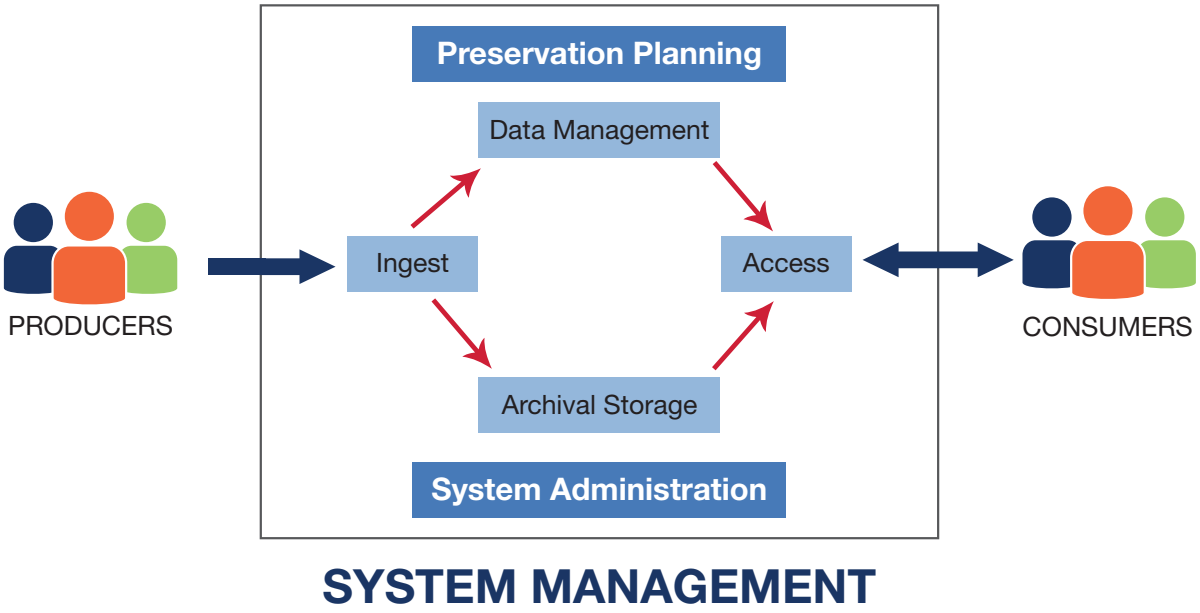
Traditional lab notebooks were paper-based without much beyond written notes and diagrams. Now, data, metadata, analysis and annotations are all captured in e-lab notebooks. SDB, for example, allows users not only to store notebook entries but also to store the raw data and processed data that was used to create that notebook entry, and link everything together, and have them accessible in a long term archive.

Today, repurposing content and data for new use is increasingly important as the pace and size of the flood of experimental data overwhelms efforts to analyze it. Failure to effectively preserve such a valuable resource undermines the opportunity to review the content in the light of new tools and accumulating insights made by others.

The life science world is steadily taking notice. The National Science Foundation, for example, is making a big push in academia to require Data Management Plans be included in research proposals to secure funding. NSF wants funding applicants to demonstrate that that there is a plan in place to provide long-term digital curation for the outputs of any proposed research.

The emerging Era of Big Data has stirred interest in digital archiving across virtually all industry segments; nowhere is digital archiving more needed than in biomedical research and healthcare where the tools and their output are being dramatically transformed.

Deploying a comprehensive digital archive aligned around important research and business goals can produce important benefits, reduce operational costs, and greatly facilitate regulatory compliance. For more information about Tessella consulting capacities in long-term archiving or the Tessella Safety Deposit Box platform, visit http://www.digital-preservation.com



OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

The OAIS reference model describes the set of core functional capabilities and an information model that should be implemented in a digital archiving solution. The information model is based on a series of "Information Packages" that contain all of the necessary components including content and metadata for a user to understand and have the means to access the information being conveyed in digital form. For more information: http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

SAFE PROMOTES PKI USE

Currently, digital signature technology relies on a PKI – public key infrastructure - and third party certification authorities are required to attest to the validity of signature certificates. The SAFE Biopharma Standard was created to support these efforts.

> "But just as software comes and goes, so do certification bodies," says Evans. "If you are using digital signatures, you must think about putting in place a process to authenticate that digital signature as soon as it comes into the archive. Then you can say the signature was valid and audited when it arrived and you can demonstrate it has not changed since."

i. *DNA Sequencing Caught in Deluge of Data,* New York Times, Nov. 30, 2011, http://www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html?_r=1&ref=science

ii. *ISO14721:2003*; http://public.ccsds.org/publications/archive/650x0b1.PDF

iii. *The Long-Term Preservation of Digital Information,* http://www.digital-preservation.com/wp-content/uploads/DigitalArchiving.pdf

iv. *Tessella Digital Preservation,* http://www.digital-preservation.com/