The eArchive Service: A practical, cost-effective and sustainable approach for preserving authentic electronic records at AstraZeneca

Mark Pickering; AstraZeneca Archives & Records Management; Wilmington, DE, US. Mark Evans; Tessella Inc; Rockville, MD, US

Abstract

Long-term retention of authentic electronic records is of significant importance to the pharmaceutical industry, from many standpoints including legal, regulatory and intellectual property protection. The risk of losing access to critical business information in the future could potentially be detrimental to a pharmaceutical business. The industry in the main has relied on a variety of archival approaches, such as using file systems and content management systems and keeping operational systems alive beyond their intended lifetime; unfortunately, none of these approaches are sustainable in the long-term and do not address the serious issue of loss of access through technology obsolescence.

The eArchive service has been developed by AstraZeneca to manage the long-term preservation of electronic records, and to begin to address the issue of technology obsolescence. The system is intended to ingest and reliably maintain electronic records from all areas of the global business with retention periods ranging from seven years to many decades. A wide variety of both proprietary and non-proprietary formats are used for the creation of AstraZeneca records; to date, there has been little focus on generating records in formats suitable for long-term preservation. As a result, many records requiring long-term retention may become at risk of format obsolescence unless action is taken. The eArchive service contains mechanisms to mitigate against such risks and facilitate continued access to authentic content throughout a records retention period. In addition to ensuring continued access, strict records management policies must also be enforced. Material must be disposed of once its retention period has been reached, as long as it is not required for ongoing litigation (on legal hold).

This paper describes a practical and extensible approach employed by the eArchive service for providing cost-effective and sustainable access to electronic records over time. The solution has been developed in accordance with the Open Archival Information System (OAIS) reference model and AstraZeneca's own enterprise architecture. Consideration has also been given to records management activities, such as appraisal and disposition that are not described within the OAIS reference model.

The primary approach has been to maximize the use of robust COTS (Commercial off the Shelf) products to form the core of the solution. Customizations are applied where appropriate. Concepts and tools developed by the EU Planets program have also been incorporated into the design, as well as aspects of metadata standards such as PREMIS and Dublin Core.

Format migration has been adopted as the long-term preservation mechanism, in addition to a policy of accepting only a limited number of "Preservation Ready" formats. As a result, many of the format migrations will take place prior to ingest into the eArchive system. This paper will describe in detail the approach to preservation policy, identification and selection of preservation pathways and methods of validating the authenticity for long-term preservation of digital images. The tiered fidelity model, based on business requirements, will also be introduced and described in detail.

The Situation

Business and Industry Drivers

Increasingly challenging legal and regulatory requirements for keeping company records, coupled with the impact of the information explosion of the last 25 years, have presented a great challenge to the pharmaceutical industry to gain control over the exponential growth in their electronic records. Cost containment continues to be a key industry driver, as worldwide pharmaceutical industry revenue growth, while still positive, is slowing. This is due to pressure on healthcare costs, exacerbated by the current economic downturn, as well as increased competition from generic medicines. Cost pressures are likely to continue, especially in the US.

AstraZeneca's continued drive for efficiency and effectiveness, in response to competitive pressures, demands attention to containment of IS costs. Decommissioning of obsolete systems reduces system management and storage costs; this is the cornerstone of the financial case to deliver an enterprise electronic archive capability (eArchive) at AstraZeneca. *Providing a positive Return on Investment within five years was an absolute requirement for gaining approval to develop the eArchive.*

Failure to comply with regulatory, legal and internal information requirements would be potentially detrimental to AstraZeneca's reputation and competitive position; hence, compliance is a key driver for accurate retention and destruction of electronic records. *Key records are retained from decommissioned systems and disposed in a timely fashion, in order to meet retention and compliance requirements.*

Record Lifecycle and Retention

The full Record Lifecycle of an electronic record includes the Creation, Use and Maintenance stages, followed by Disposition. During the Maintenance stage, less-active records that require long-term retention may be archived to ensure preservation, safekeeping, and accurate disposition of the records. The eArchive focuses on records that have a retention requirement of seven years or greater, and are no longer actively used in the operational environment. Required retention periods at AstraZeneca can be many decades, and retention of some records is permanent. Cost-effective management of this duration requires infrequent migration of records to standards that meet the test of time, preferably enabling holding periods of ten years or more between migrations.

AstraZeneca's Electronic Records

AstraZeneca's electronic records include a broad range of information types, including office documents, business transactions, images, engineering drawings, audio/video files, instrument files, laboratory notebooks, and many more with a mix of proprietary and published formats. The records represent business evidence necessary to prove compliance, intellectual property, the drug development process and business transactions. Some records are created in formats suitable for long-term preservation; others are not. Process improvement is underway at AstraZeneca to drive the design of records that are "born- ready" for preservation into new systems; however, there will be a long legacy of electronic records that will require transformation in order to ensure preservation.

Where are we now, and what is the vision?

The eArchive Programme and other programmes underway within AstraZeneca Archives & Records Management continue to address the dual drivers of cost containment and compliance. As a result, AstraZeneca is on the leading edge among pharmaceutical companies in the delivery of enterprise-wide capabilities for the management of electronic records in a cost-effective and compliant manner.

The vision of the eArchive Service speaks to the future of the long-term management of electronic records at AstraZeneca:

> *"Deliver a process-driven enterprise-wide service, embraced by the AstraZeneca community, that provides long-term preservation of electronic records as an integral part of the AstraZeneca information lifecycle."*

The Solution

Introduction to the Solution

The eArchive Service is a combination of processes and technology developed to meet AstraZeneca's electronic archiving needs.

Solution = (Appraisal +
Ingest Pipeline +
Ongoing Management +
Retrieval +
Disposition) +
Technology

The business customer collects key information by performing a self-assessment that precedes the appraisal. This provides an initial test to see if their records meet the basic criteria for entry into the eArchive. The appraisal includes a detailed review of the records proposed for archiving, including classification of records, assignment of retention according to the classification, technical review of record formats, proposal of an extraction and transformation process, discussion of metadata and other preliminary steps that must precede archiving.

The Ingest Pipeline includes many key steps for preparing and ingesting records into the eArchive. These steps include:

- Record Preparation and Transformation – ensuring that records are in a preservation-ready format
- Metadata Mapping – reconciliation of record metadata with the eArchive Metadata Model
- Building the SIP - the Submission Information Package (SIP) is created containing all information to be archived
- Transfer – the SIP is securely transferred to eArchive Service location
- Ingest – the electronic records are ingested into the eArchive

Ongoing Management of the records contained within the eArchive includes preservation activities, storage and media management, migration activities, continual record integrity checking and a host of other management activities.

Retrieval of records is provided from within the eArchive for records that can be displayed through standard desktop viewers. Export can be performed as needed to meet other customer needs.

Disposition is performed using standard functionality of the application software in conjunction with AstraZeneca's Global Retention and Disposition policy and Legal Hold Process.

The technology that supports the eArchive Service includes an Enterprise Content Management system that is widely used at AstraZeneca and other pharmaceutical companies, plus add-on Record Management components. Also, a number of COTS and OpenSource tools are used to perform transformations and fidelity checks. The solution is mounted on commodity Unix and Wintel servers in a World-Class Datacenter in Sweden with data replication to a Disaster Recovery site in the UK.

Alignment with and extension of OAIS

The eArchive solution has been architected and designed in alignment with the well-adopted Open Archival Information System (OAIS) reference model [1] from both a functional and informational perspective. In particular, some aspects of the

"Preservation Planning" component have been incorporated into the design to support continued access to authentic content throughout the retention period. As described above, the eArchive service must provide a comprehensive set of records management functionality, in addition to acting as an archive for electronic records. As a result, the OAIS functional components have been extended to take these broader requirements into account.

The principle extensions to the OAIS reference model include appraisal, retention management and legal hold management. As described below, the preservation policy enforced by the eArchive service is to manage only a small set of "Preservation Ready" formats. As a result, a set of pre-ingest services are provided to facilitate the transformation of source content to a "Preservation Ready" format.

The diagram below illustrates the main functional components in the eArchive architecture:

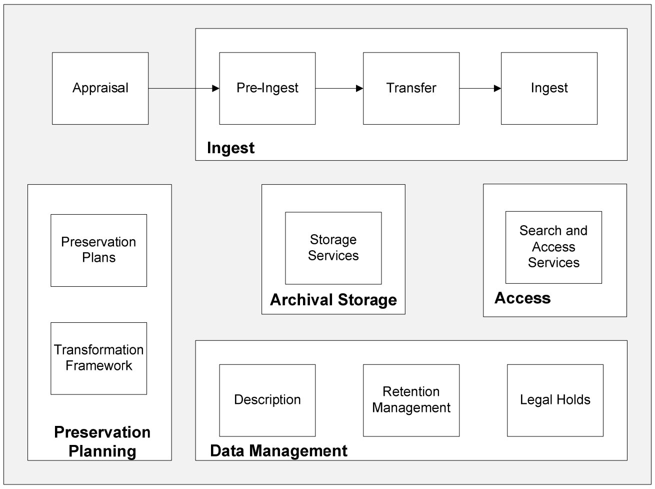

***Figure 1**. eArchive high level components*

Informational Model

The OAIS concept of an information package is represented through the eArchive service at a logical level. The physical manifestation of an information package is not a single encapsulated entity; rather, it is a series of digital objects that are tightly related through metadata. In addition to the electronic record content, a SIP created by the business area contains descriptive and basic technical metadata. As the SIP passes through the ingest pipeline, additional metadata is added and an Archival Information Package (AIP) is created and placed into the storage repository. Users that perform searches and require access to archival content can extract Dissemination Information Packages (DIP) that can contain one or more records.

An electronic record is a highly conceptual construct that can have a complex relationship with digital files. In the simple case, a record can be represented by a single file; in more complex cases, a record can be represented by multiple files or even by parts of multiple files. A preservation action may also create a new set of digital files that may have a different set of relationships than the original. In order to support this complex relationship, the eArchive service employs data constructs taken from both the European funded Planets project [2] and the PREMIS metadata standard [3]. The Planets concept of a Manifestation (Representation in PREMIS) is used to bridge the relationship between records and files. A Manifestation represents a single view (or rendition) of a record; a record may have multiple manifestations over time. For example, a second manifestation may be created to support efficient dissemination (a presentation manifestation) or mitigate against format obsolescence (a preservation manifestation). The high level logical data entities and relationships are shown in the diagram below:

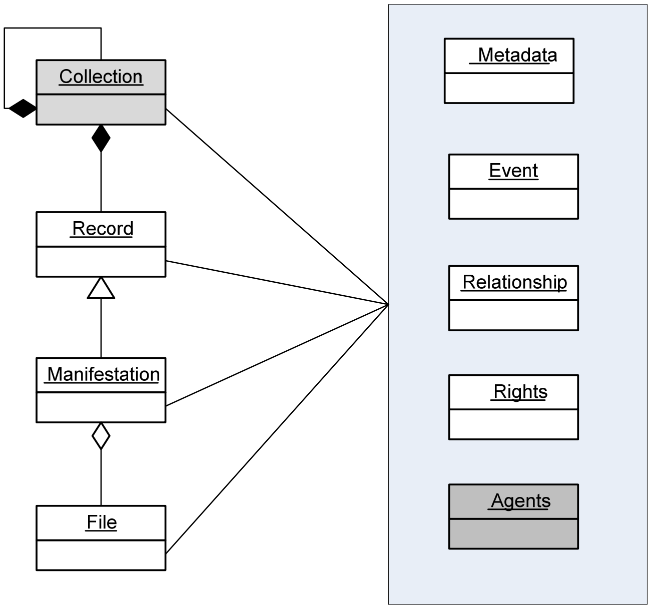

***Figure 2.** eArchive Logical data entities.*

Grayed-out entities are reserved for a future release of the eArchive service. Most of the entities on the right of Figure 2 are taken from the PREMIS model and can be captured for all of the entities on the left.

As illustrated above, metadata can be captured for all entities in the logical data model and an extensible scheme has been created that supports this. In defining such a scheme, many standards were considered, including Dublin Core [4], METS [5], PREMIS and the enterprise model that is used throughout the AstraZeneca business. No single standard could provide complete coverage for all of the metadata requirements, so key aspects have been taken from many standards. As examples, many of the PREMIS semantic units have been adopted for file level metadata, and Dublin core elements have been adopted for descriptive metadata.

Descriptive metadata is captured at the collection and record entities and divided into five broad categories:

- Core Metadata – Elements that are common to all records across the business; mostly mandatory and derived from Dublin Core
- Enterprise Metadata – An extension to core metadata to provide more business related metadata
- Record Type Metadata – Metadata elements used by a particular area of the business, such as R&D and Sales and Marketing
- Extended Metadata – Any additional elements that do not fit into one of the other categories

File level metadata includes the following:

- Original name and relative path so that any required hierarchy can be reproduced
- Format information – Includes a PRONOM identifier [6] and the degree of success of format identification during the ingest pipeline
- Fixity – A checksum and details of the algorithm used; the checksum of the file is periodically validated to ensure the continued integrity of files
- Details of any electronic signature that may have been applied to the file

eArchive approach to Preservation

The approach to long-term preservation employed within the eArchive service has been influenced by many factors including:

- Current best practice and standards in the domain
- Operational sustainability
- Level of technical complexity
- Current business environment
- Corporate policy

The eArchive Service provides both bit level preservation capabilities (through the use of file level integrity checksums) and file format migration and validation to guard against file format obsolescence. In addition, a policy has been set to only accept a small number of “Preservation Ready” file formats for ingestion. A “Preservation Ready” format is one that is deemed to exhibit most, if not all, of the following qualities:

- **Openness:** existence of format standards (published documentation and open disclosure by international standards).
- **High Adoption:** widespread popularity and use (both throughout the industry and within the business).
- **Self-Documenting:** The inclusion of descriptive and technical metadata for self-describing the content.
- **Suitability:** Ability to represent a wide range of records created by the business; includes the ability to provide an authentic representation of the original record with respect to content, structure, behavior and appearance; must be able to represent the semantics and complexity of the information to be preserved.

The file format selection process was influenced by a number of previous studies including those conducted by national archives and libraries [7],[8], [9] ,[10] and industry wide organizations [11],[12].

By adopting the approach of only accepting information in few and standardized formats, the long-term archival management of a substantial collection of records is somewhat simplified. Over time, this approach reduces the overall requirement to perform record migrations to enable the continued access to archived records. The same policy can also be used to direct the future development of new operational systems, so that they create content in a “Preservation Ready” format. A list of the baseline set of “Preservation Ready” formats for a variety of content types is provided in the table below, however it is expected that this list will be dynamic over time.

***Table 1**. Initial eArchive “Preservation Ready” formats*

Content type	Accepted Formats / Encodings
Data Centric	XML, UTF
Document Centric	OOXML, PDF, PDF/A, HTML
Rich Media: Images	TIFF, PNG, JPEG2000
Rich Media : Audio	MP3, WAV
Rich Media: Video	MPEG-4 Part 10

Transformation Framework

As mentioned, to date, there has been no control on the use of file formats for electronic records within the business. Current operational systems are creating content in a wide number of file formats, many of which are not in the list of acceptable preservation formats. In order to meet the preservation policy described above, such content will require transformation from its original format to one of those described in the table above. In addition, it is conceivable that one of the baseline “Preservation Ready” formats may become obsolete within the retention period of content stored in that particular format.

The eArchive service contains a transformation framework that provides an extensible approach to tackling the requirement to perform file format transformations over time. The framework provides:

- Tools to perform the file format transformation
- Tools to validate the authenticity of a transformation
- A set of available preservation pathways
- Extensibility to add new tools in the future

It is envisioned that the framework over time will support a fully automated transformation and validation process and significantly reduce the need for manual inspection of transformations.

A preservation pathway is defined as an available transformation route from a source format to a “Preservation Ready” format using a specified tool in a specified configuration. An example of preservation pathways for images is described later in this paper. A baseline set of pathways has been identified based on the current use of legacy formats throughout the business and available tools for transformation.

An essential component of the transformation framework is the ability to validate a successful transformation, referred to as "Fidelity Testing." Two approaches have been adopted in providing this functional capability:

- Property extraction and comparison
- Content comparison

In the first method, a set of "essential characteristics" (defining the essence of the content that should endure over time) has to be established for a set of content. These determine which properties are then extracted and subsequently compared pre- and post-transformation. The notion of essential characteristics is relatively new and much work is still required to reach a better understanding and identify best practice. The Planets Plato tool [13] is one example of recent research in this field; however, for some content types, a solid set of essential characteristics can be determined. An example for images is given below.

***Table 2**. Example of essential characteristics for images*

Essential Characteristic
Dimensions (height , width)
Aspect ratio
Resolution
Number of colors

The extraction and comparison of properties, before and after transformation, provides some degree of confidence that authenticity has been maintained. However it does not give any real indication that the actual content has been maintained. For example, the resolution of an image may be preserved, but the color of an individual pixel or set of pixels may have changed. There is a need then to supplement property extraction tools with a mechanism based on content comparison to further validate and demonstrate that authenticity has been maintained. Content comparison tools are used in the eArchive service for the validation of image transformations and certain document transformations. As with the case of essential characteristics, this is an emerging area, and as more tools become available, they can be evaluated and integrated into the eArchive.

Image Management Toolset

As illustrated in Table 1 above, several image formats are accepted by the eArchive service to meet a variety of requirements. As a longtime standard, TIFF has been selected to provide continuity with many existing images, especially where there is reluctance to transform. JPEG 2000 has been selected as the premier image preservation format, supporting several specifications of images from lossless (to support high-quality scientific images) to high compression (see Tiered Fidelity Model later in this section). PNG provides a high-quality lossless standard with superior compression (up to 99%) for vector graphics and other images of lower complexity.

The toolset selected for image transformation and validation is a mature, full-feature OpenSource tool that operates in command line mode. The tool supports a wide variety of image transformations, including all of the currently supported input and output format types required by the eArchive. Quality-checking capabilities are also provided by the toolset, enabling comparison of a lossless output image with the original. The toolset is operated through calls to the transformation framework, and provides suitable performance to meet AstraZeneca's current needs. Except for process initiation and error resolution, the entire transformation and fidelity checking process operates in unattended mode, and can support transformation of thousands of images in a single execution.

To simplify appraisal and transformation of images, a Tiered Fidelity Model has been adopted. The model provides high- and standard-resolution specifications that are selected as transformation targets during appraisal. This approach helps avoid the time-consuming trap of selecting from a potentially endless array of output image specifications. As an example, TIFF images whose value lies in the details of color and shape (such as tissue samples) might be transformed to a mathematically lossless JPEG 2000 image (with a space savings of approximately 50%); while a photograph of the CEO might be transformed to a standard specification that provides suitable quality for the subject matter, but with greater compression levels (with space savings of 95% +) as well as higher loss. The selection of a target resolution is based strictly on business needs; selection of a high-resolution specification requires a suitable business case.

Process and Change Management

Although this paper has focused on the preservation and technical aspects of the eArchive, the process and change management aspects of the eArchive Programme have received equal focus and manpower in order to bring the eArchive Service to successful delivery.

The business processes, rather than the supporting technology, are considered to be the driving force of the eArchive Service. Without well-defined processes, the service would depend strictly on the knowledge and capabilities of individuals, which could change over time depending on who is involved in service delivery. The establishment of well-defined business processes clearly spells out the key activities, as well as the roles and responsibilities of all participants. Consistently good service would be difficult to deliver without a high-quality process framework.

Change management has also been a key aspect of delivery of the eArchive. In order to secure funding for the service, support and approval of a wide range of business people and executives has been required. An active Business Reference Team has provided input, guidance and access to key business staff since the beginning of the business case. In addition, a Sponsor Group has provided essential support at the executive level. In order to gain buy-in for an enterprise-wide capability, it is essential to "bring people along for the journey." Without an appropriate change management effort, it is difficult to gain the trust of the business. In many cases, electronic records are cherished artifacts of business and scientific endeavors; without trust that the eArchive Service will suitably preserve and manage the documents, reluctance to part with records may cause under-adoption of the service.

References

[1] Consultative Committee for Space Data Systems (2002). Reference Model for an Open Archival Information System (OAIS). CCDS 650.0-B-1, Blue Book http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] Planets Project – http://planets-project.eu

[3] PREMIS Data Dictionary for Preservation Metadata v2. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[4] Dublin Core Metadata Initiative. http://dublincore.org

[5] Library of Congress Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/

[6] PRONOM Technical Registry http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[7] Brown, Adrian: Digital Preservation Guidance Note 1: Selecting File Formats for Long-Term Preservation, UK National Archives 19th June 2003.

[8] Rog, Judith, Wijk, Caroline van: Evaluating File Formats for Long-term Preservation, National Library of the Netherlands

[9] Library of Congress: Sustainability of Digital Formats – Planning for Library of Congress Collections

[10] Guidelines for Computer File Types – Library and Archives of Canada http://www.collectionscanada.gc.ca/government/products-services/007002-3017-e.html

[11] SNIA activity on preservation formats - http://www.snia.org/forums/dmf/programs/ltacsi/100_year/SNIA-DMF_Towards_%20SD-SCDF_20070412.pdf

[12] Formats For Digital Preservation: A Review of Alternatives and Issues - http://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf

[13] Planets Plato tool http://www.ifs.tuwien.ac.at/dp/plato/intro.html

Author Biography

***Mark Pickering** has had a long career in IS/IT following studies at Pennsylvania State University (BS 1977), Drexel University and West Chester University. His work has focused on the pragmatic integration of systems and software to deliver high-performance enterprise class systems at US financial institutions, Eastman Kodak and AstraZeneca. Recent work includes development of capabilities for long-term preservation and archiving of electronic records.*

***Mark Evans** is the Digital Archiving Practice Manager for Tessella Inc, and provides oversight for all of Tessella's archiving projects in North America. Mark has been involved in the specification, architecture and design of digital archiving solutions for the past 8 years, and has worked on several national programs. He was part of the Tessella team that developed the world's first national digital archive for the UK National Archives in 2003. More recently he has played a major role on the NARA Electronic Records Archive (ERA) program, and is currently leading the infrastructure development on an NSF DataNet project team, led by John Hopkins University.*