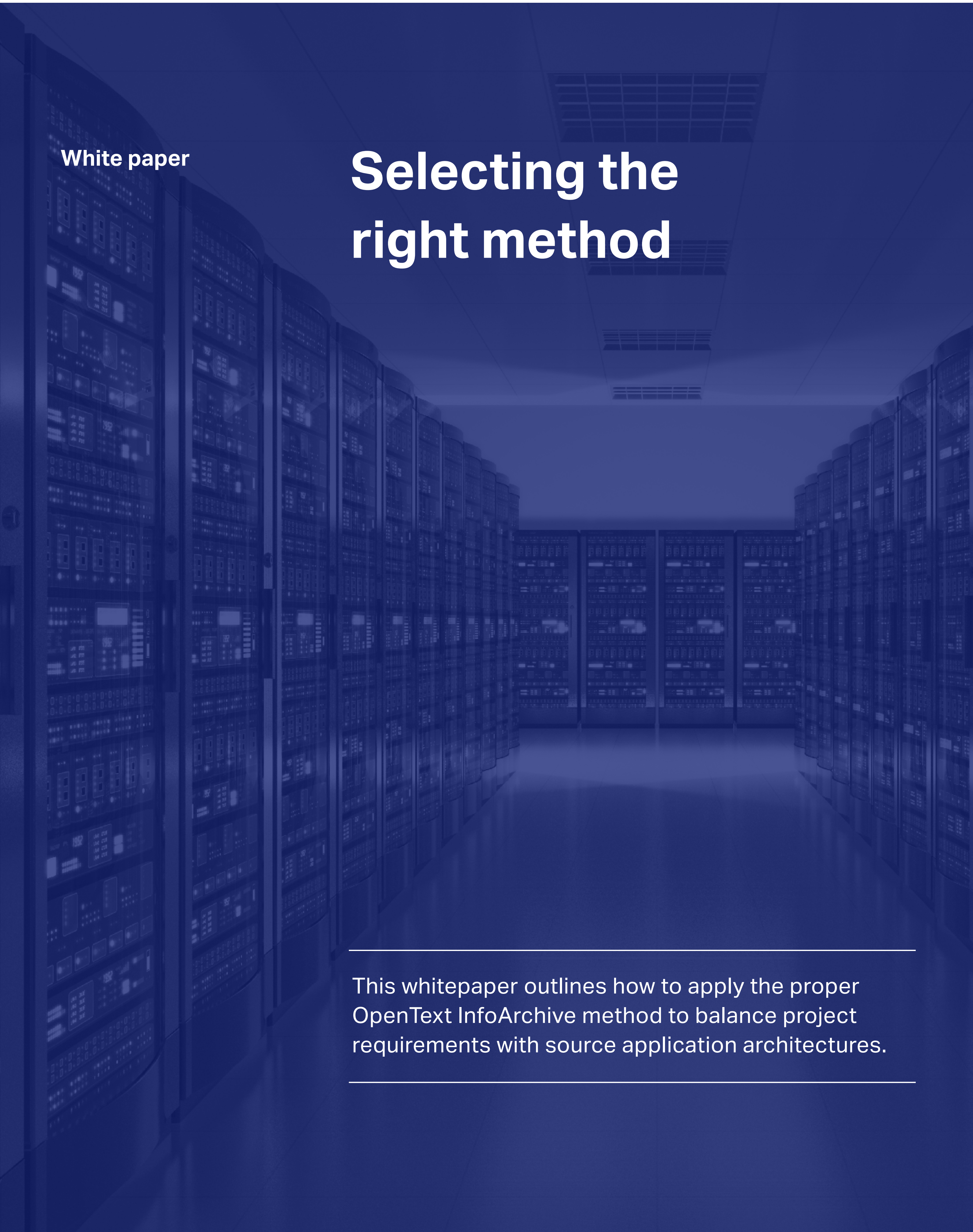


White paper

Selecting the right method

This whitepaper outlines how to apply the proper OpenText InfoArchive method to balance project requirements with source application architectures.

Contents

OpenText™ InfoArchive is an information management and archiving solution that supports different enterprise needs for ingesting application data of all types. It provides four methods of ingestion to meet the needs of competing project requirements and optimize the environment to the source application. With InfoArchive, there is no need to take a one-size-fits-all approach. This white paper provides a guide to help organizations choose which of the four InfoArchive methods for preserving and reusing information is right for each of their applications.

The four methods

Step 1: Setting archiving goals

InfoArchive is an integrated product suite designed for managing and archiving application information. It supports three core use cases based on short and long-term archiving goals:

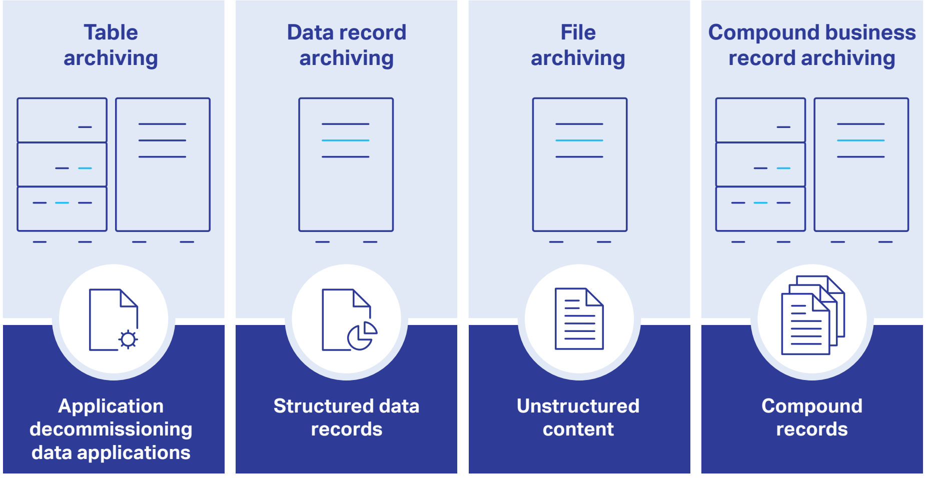
- **Cost take-out.** InfoArchive can provide a repository for data from legacy and redundant applications that might have been superseded by an ERP system, replicated during an acquisition or must be decommissioned as part of a business sale, closure or industry mandate. Data from applications that a company migrates to InfoArchive will remain accessible for business reporting, audits or compliance with data retention regulations. Meanwhile, the organization can shut down the applications and save all the costs associated with supporting and maintaining them.
- **Optimize.** Companies can also use InfoArchive for periodic archiving of data and content from "live" business applications to reduce costs for production environments, enable compliant data retention and optimize application and infrastructure performance. In addition to savings on storage costs, companies can reduce costs for backups, system administration, servers and database licensing costs.
- **IT transformation and reuse.** Increasingly, organizations require access to information for new and innovative uses. Advanced and predictive analytics, as well as application modernization programs, top the list of major programs that demand fresh approaches to information access. InfoArchive supports these requirements by serving as a platform for data aggregation and management that offers access to business records in bulk via the Hadoop® Distributed File System (HDFS) or individually to other business applications via web services.

InfoArchive uses extensible mark-up language (XML) as the format for preserving data and metadata for long term, platform-independent retention. Data/metadata from multiple sources can be aggregated and represented as XML files. This representation can actually translate to a business object.

The technology archives the native XML data and structure, allowing users to query content efficiently and accurately, at any level of detail. Users can also transform that data into views formatted for print, websites, mobile devices and other channels. In addition, user interface development tools use declarative XML syntax. This greatly reduces the need for, and cost of, custom programming to deliver interactive content.

With InfoArchive, companies can manage an unlimited number of data structures in a single repository. They can store both structured data and unstructured content in a single record and access all the information they need for their business processes and reporting from a single query. InfoArchive is the only solution that delivers all of the following methods to ingest information for archiving:

- Table archiving
- Data record archiving
- File archiving
- Compound record archiving



Organizations can use any of these options to decommission applications and archive active applications. Such flexibility is particularly valuable in decommissioning programs that involve many applications and information formats. Active archiving may use all of the options except table archiving. When information aggregation and reuse is key, data record, file and compound record archiving options should be considered.

Step 2: Understanding the source application

When choosing the best archiving method for a particular application, a business must consider two key questions:

- What type of information is being archived?
- How will users access that information going forward? Most information is managed in one of the following systems:
  - Transaction systems
  - Print stream systems
  - Content and image repositories
  - Interaction systems
  - Collaborative systems

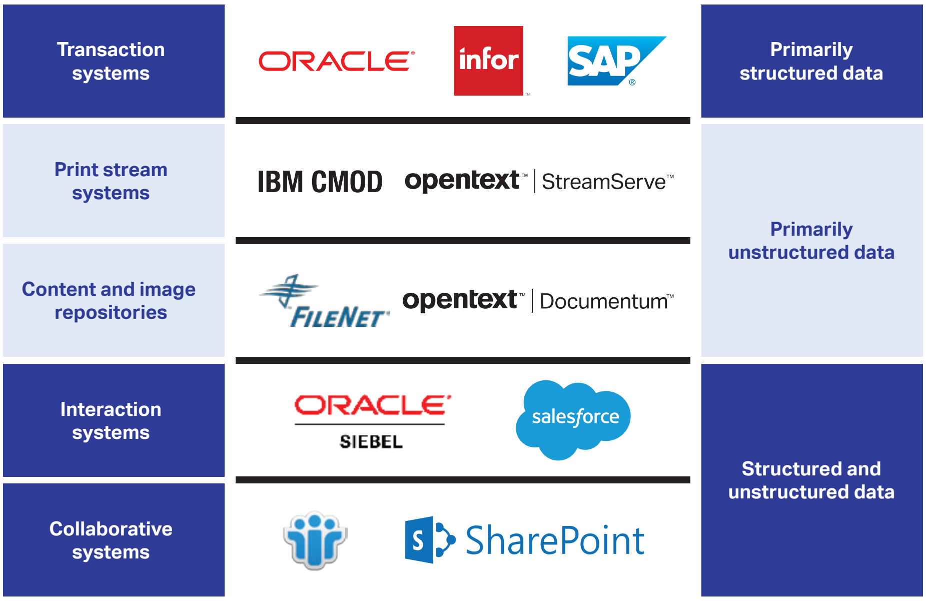


Figure 1: Key systems for information management

Transaction systems

Transaction systems have databases that hold details of past business events (such as those related to processes for accounting, ERP, enterprise asset management or supply chain management). They are used to maintain reference data in master files, record activities in transaction files and store old records in transaction history tables. They may include cloud-based systems and allow many people to add small bits of detail over time.

Print stream systems

Traditionally referred to as COLD (computer output to laser disk) systems, these systems store print stream information for long-term preservation. Most of this information involves customer communications. Other examples include greenbar reporting systems.

Content and image repositories

Content and image repositories store unstructured information and metadata, typically in their native formats. Examples include the traditional Enterprise Content Management systems, as well as storage-based systems.

Interaction systems

Interaction systems connect users with an organization for quick access to complete information. Examples include systems that support customer relationship management (CRM) and collaborative tasks. These systems include data, as well as transaction, grouping and unstructured files.

Collaborative systems

Collaborative systems allow groups of individuals to share information and communicate with each other around specific topics. These systems have all the characteristics of interaction systems but generally cater to a less structured approach. Notable examples include OpenText™ Documentum™ eRoom™, Microsoft® SharePoint® and IBM® Notes®.

Step 3: Selecting the appropriate archiving method

InfoArchive offers four archiving methods that are optimized based on the format of data or content being archived, the ease of extraction and up-front analysis and how the information is to be used after it is moved to the InfoArchive repository. Having this choice is a critical success factor for large-scale information management programs that involve a wide range of applications.

Table archiving

- Archives application tables (with content)
- Data-rich applications
- Reuses existing ETL tools and standards
- Reduces application analysis to enable rapid decommissioning and rapid time to value
- Offers limited reusability

Table archiving is a method that models structured information in the source application as XML in InfoArchive—table for table, column for column and row for row. This method lets organizations quickly decommission applications, providing a fast time to value, while preserving all data relationships for future queries and reports.

Organizations can use table archiving to migrate structured data in application tables and linked files from transaction systems to InfoArchive with few, if any transformations. Table archiving can reduce the up-front analysis involved with decommissioning an application and virtually eliminate the data integrity risks associated with other archiving methods. Because information is stored in an aggregated manner, access is less flexible than the record-based methods described below. Access is traditionally limited to query-based (list) reports.

Table archiving is used primarily with transaction systems that contain structured data and linked files, as well as with some interaction systems for the purpose of application decommissioning. Information preserved using the table archiving method may be reprocessed later into record formats for reuse scenarios.

For additional information on table archiving, see the Application Decommissioning Solution Package.

Data archiving

- Archives individual data records (with linked content files)
- Suits data-rich applications
- Enables data integration (removes application silos)
- Supports data extension (provides new context)
- Ideal for reuse

Data archiving involves identifying entities within the target application, then extracting and aggregating the associated data into a single XML-based record. Archiving structured data in XML files provides a future-proof format that is ideally suited to long-term retention, access and comprehension. Any data structure can be modelled as XML, and InfoArchive does not impose its schema. As such, one record may contain multiple XML packets and information from multiple systems may be drawn together according to the requirements of the project—all while preserving a complex multi-system chain of custody.

Data archiving is especially useful when companies want to reuse the information, while reducing costs, in a context that is different from the source application. Additionally, data archiving is well suited for active archiving of live systems. It typically requires more business-oriented analysis of application data than table archiving does. Examples of appropriate data archiving use are SWIFT transactions, sales histories and patient histories.

This method has two important advantages. First, the complex data model of the source application is transformed into a simple data model in the archive. This can reduce costs and simplify future access. Second, because there is no direct link between the source application and data in the archive, any change in the source application does not force a change in the archive. When a change in the source application results in an update to the archive data model, InfoArchive ensures that results for searches of data sets include all records across all the changed data.

Data archiving also helps organizations create new value for their data. By extending or recording metadata, they can harmonize records and support searching and filtering across data sets. Data archiving is used with transaction systems for active archiving of individual structured data records (for example, transaction history tables). It is also used with interaction systems (for decommissioning data and queries, optimizing searches and advanced analytics) and with content systems. Because it presents data as single records, it is ideal for archiving information according to government requirements and legal mandates.

File archiving

- Archives any type of unstructured content
- Suits content oriented applications
- Enables content integration
- Ideal for reuse

File archiving involves archiving unstructured data and its associated metadata into a single record. The information can be preserved in original format and/or transformed into a more future-proof format, such as PDF-A. One record may contain multiple files to create sets of related information.

Metadata is particularly important when working with the file archiving method. Attributes may be derived from the content itself or associated with other systems.

File archiving is especially useful when organizations want to reuse the information in a context that is different from the source application. Print streams and media archives are an excellent example of file archiving. Transforming large print streams (such as customer statements or explanations of benefits) from print-oriented formats (such as AFP or metacode) into a PDF for presentation on company web platforms can improve ROI and customer satisfaction. Image archives (typically multi-page TIF with limited attributes) can be reprocessed to add document type and even full-text OCR (optical character recognition).

File archiving can be used with transaction, content or reporting systems to decommission content, optimize data search and retrieval and simplify user access to information across systems. Businesses can use this method to archive any kind of unstructured content files and metadata, including print streams. It extracts value from archived content by addressing files via the associated application metadata, rather than directly from the infrastructure.

Pushing files into InfoArchive the moment they become "inactive" can reduce costs and increase system performance. Having all of an organization's "inactive" files in the same archiving platform, can also improve data discovery and enhance the options for reusing content and data through new applications or web browsers. Please note the information can still be accessed via end-user portals with high availability.

Compound-record archiving

- Combines structured data and unstructured content elements to create a business record
- Any application or combining data/content from multiple applications
- Captures and retains the business context at a point in time
- Ideal for archiving complex records like financial trades, case files or laboratory testing records

As the name implies, compound-record archiving preserves structured and unstructured data into a single record. The structured elements are modelled as XML and the unstructured elements may be preserved in the original format or transformed into a more future-proof format, such as PDF-A. One record will contain multiple files of related information.

Compound-record archiving provides the only mechanism that can compliantly archive systems with a blend of structured information (such as wikis, blogs and comments) with unstructured information (primarily attachments). It serves as a proper format for decommissioning, as well as active archiving interactive systems that involve such a mix of structured data and unstructured content. Microsoft SharePoint and IBM Notes applications are notable examples.

As business processes become more complex and regulations more demanding, there is a growing need to archive business records that may contain multiple structured data and unstructured content elements that must be brought together to create the final business record, for example, financial trades, cases and laboratory reports.

With compound-record archiving, organizations can retain business events as single records and reuse them for analytics or regulatory audits. Users can search the records using pure business logic, without switching from one application to another.

A solution for structured and unstructured content

Any ERP implementation involves the creation and management of various records or business objects. Some of these records are composed of structured data that users typically enter and access through form fields. As much as 85 percent of managed information can be unstructured content, according to some estimates. This content may include digital images, video, digitally rendered faxes, email messages and text documents.

While both structured and unstructured information are usually needed to drive efficient, ERP-enabled business processes, most ERP applications do not have the robust functionality required for handling the indexing, searching, storage and security of huge volumes of unstructured information in multiple formats. A content management solution, such as OpenText™ Documentum™, is often needed to provide such support.

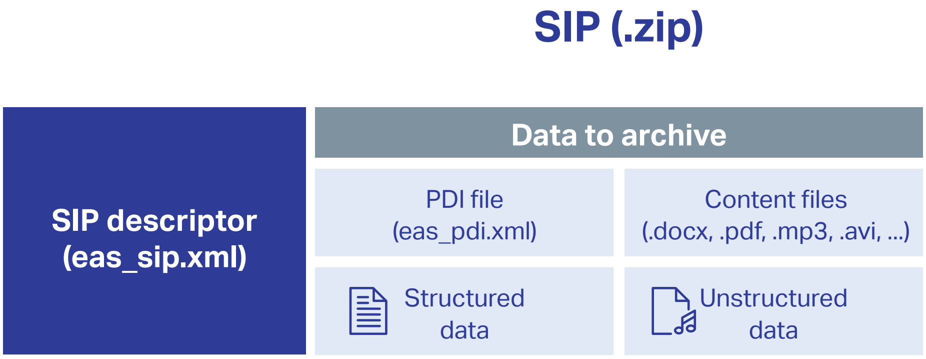
Submission information packages

With InfoArchive, information for data record, file and compound archiving is extracted by an appropriate connector and pushed to the solution in a component known in the terminology of the Open Archival Information System (OAIS) as a submission information package (SIP). The SIP is compressed as a .zip file and can be transported using any file transfer technology. The SIP is ingested to the archive and becomes an archive information package (AIP). Based on the classification, the content is stored in an archive holding. When an end user requests data from the archive, the data is delivered as a dissemination information package (DIP).

SIPs are compressed for transfer between the source application and InfoArchive to reduce network traffic. Each SIP includes:

1. **SIP metadata file**–a small XML file containing data that describes the SIP and provides data that InfoArchive uses to set retention dates, access rights and other archive policies
2. **Archive content metadata**–another XML file that holds the metadata associated with the content to be archived
3. **Content file**–the unstructured content that is to be archived

When structured data is archived to InfoArchive, the SIP does not contain any content object. The structured data that is to be archived is held as XML in the archive content metadata file and InfoArchive processes the SIP in the same way it would ingest a SIP with unstructured content. This standard process for archiving both types of content enhances efficiency and reduces total cost of ownership for InfoArchive.



SIP (.zip)

SIP descriptor (eas_sip.xml)

Data to archive

PDI file (eas_pdi.xml)

Content files (.docx, .pdf, .mp3, .avi, ...)

Structured data

Unstructured data

InfoArchive information model

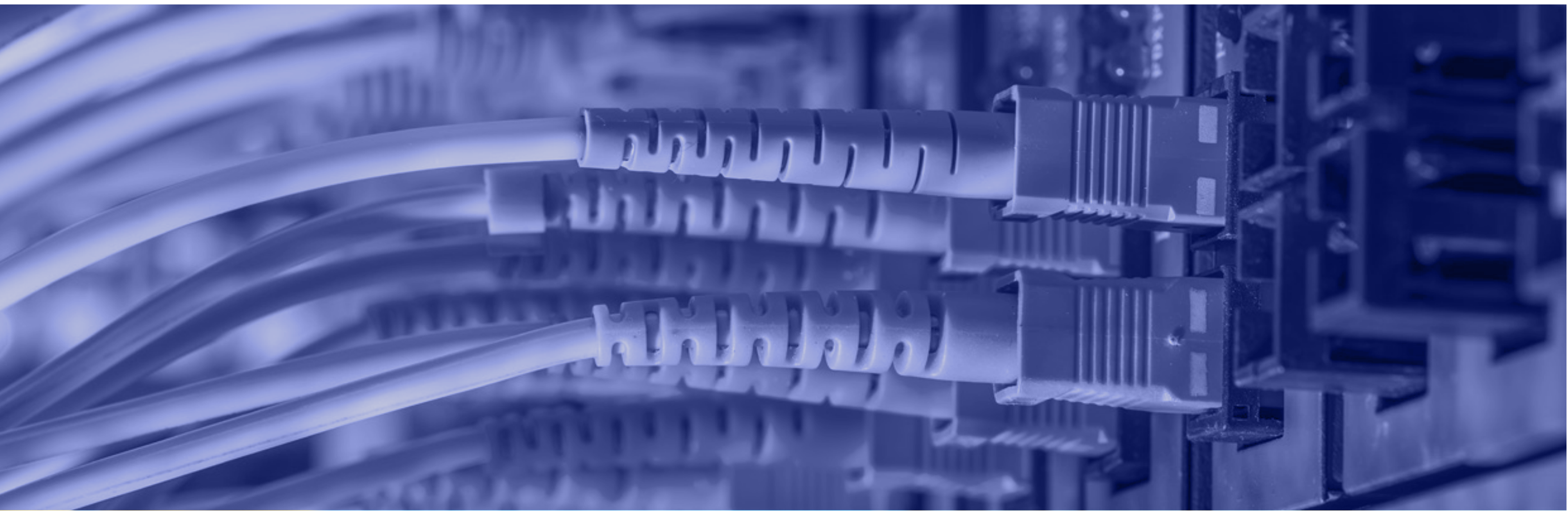
AIPs are discrete packages of information that may contain none-to-many structured data elements represented as XML and/or none-to-many unstructured data elements.

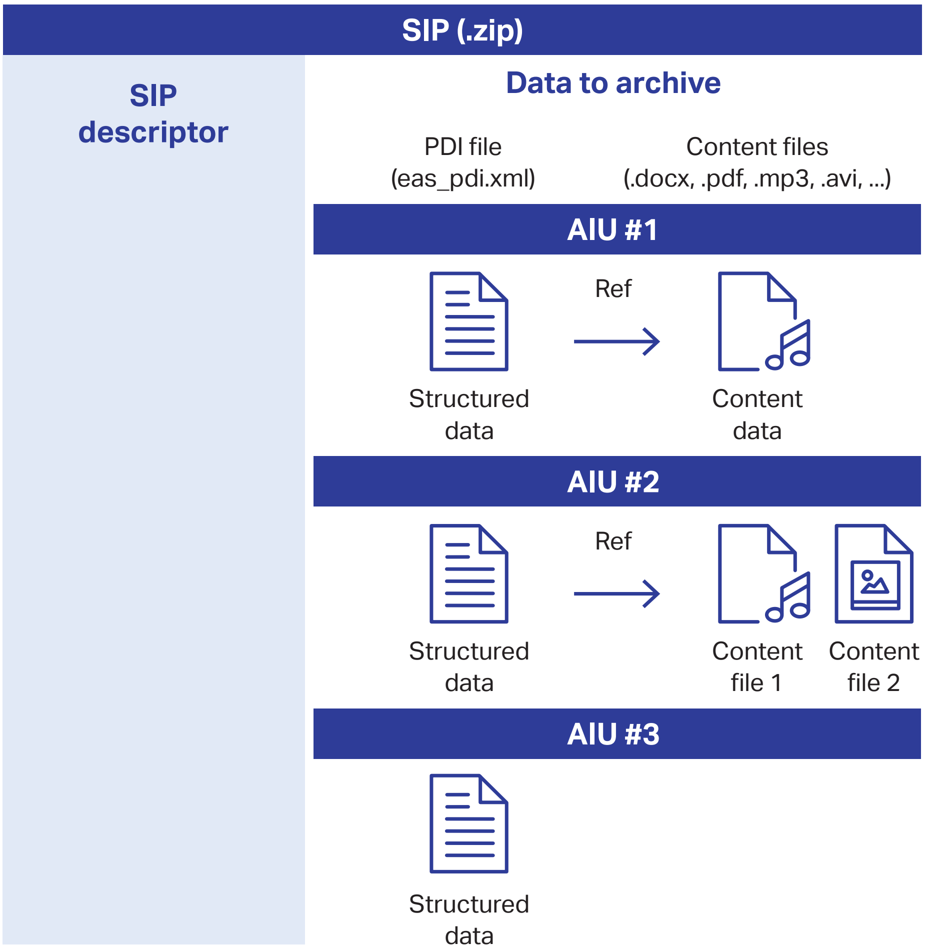
The ability to layer metadata over an AIP adds power, especially with reuse scenarios. The metadata and data elements may be extracted directly from the source application, derived from other systems or programmatically constructed. The ability to maintain separate data elements in an AIP allows InfoArchive to balance the requirements to maintain exacting standards around chain of custody with the desire to build richer data sets than existed in the original applications. For example, the raw transaction history data may be extracted, modeled as XML and verified as 100 percent accurate and complete for chain of custody purposes, while additional information from extended systems can be made part of the AIU in another data element making the AIU more usable without compromise.

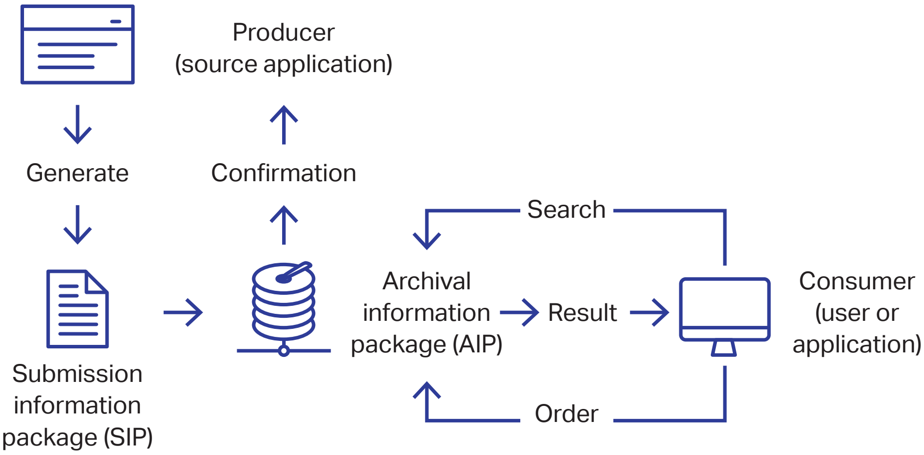
Generating SIPS

To generate a SIP, information is extracted from the source application and written to the standard SIP format for InfoArchive as an XML file. A SIP can contain one or more information records, called archive information units (AIUs). The AIUs that are extracted to the SIP are defined by rules that are part of the SIP-generation program.

When a SIP is ingested to the repository, it becomes an AIP. The AIP is stored in a logical archive-holding folder. The folder has a number of configuration options for setting management parameters that are applied to AIPs—including data retention, access control and search SLAs. AIPs extracted from different systems can be stored in the same archive holding.






Producer (source application)
Generate
Confirmation
Search
Submission information package (SIP)
Archival information package (AIP)
Result
Consumer (user or application)
Order

Bringing the concepts together

To best understand the interplay between the four InfoArchive methods and the underlying InfoArchive information model, consider the following fictitious example based on the experiences of four actual companies. GizmoRx*, a medical device manufacturer, had installed an Infor Accounts Receivable package several years ago. Hoping to increase customer satisfaction, management decided to archive invoices and make them available to its customer service department. It chose an InfoArchive product with Crawford print stream capabilities for file archiving.

After success with invoice archiving, GizmoRx decided to get more from its Infor system by archiving structured data (such as sales histories and journal entries) from its general ledger. The system became more responsive and gave the company's finance department full access to historic queries for analysis.

When GizmoRx decided to replace its aging Lawson system with a new SAP system, it found that retiring the Lawson system could help defray much of the cost for the SAP software. The company purchased additional capacity for InfoArchive and chose table archiving for the old Infor system to preserve regulated and intellectual property information for any future audits or other needs.

Six months after decommissioning the legacy application, the attorney general asked the company to produce customer lists going back 20 years, as well as a history of terms that were granted to those customers and the transaction histories. Without InfoArchive's table archiving, GizmoRX would have had to reconstruct this history manually. Instead, the company was able to analyze the table structures automatically and write a new query to satisfy the request.

GizmoRx subsequently decided to purchase more InfoArchive capacity and use file archiving services to handle archiving needs for the new SAP software. These needs include:

- Meeting regulatory requirements by purging selected data from the production environment and providing secure read-only access
- Reducing the time required for routine database maintenance, backup and disaster recovery
- Maximizing current storage and processing capacity and deferring the cost of hardware and storage upgrades

Finally, GizmoRx introduced an advanced analytics program to bind device history information from its services organization with transaction details, to better understand quality of care across its customer base. For this task, the company selected compound-record archiving for its service application and structured data archiving from the SAP transaction data. Both concepts were tied together by an additional data element in the AIPs. The AIUs were exposed as discrete elements to its Hadoop analytics platform.

** Not an actual customer*

InfoArchive is a truly versatile, all-in-one solution that can meet the full scope of information archiving, management and access requirements—whether the data is structured or unstructured, simple or highly complex. InfoArchive can handle data extracted from any source application, at any level of granularity, with the ability to scale. Regardless of the method chosen, all information is stored and managed in a compliant manner leveraging InfoArchive's comprehensive compliance features.

About OpenText

OpenText, The Information Company, enables organizations to gain insight through market leading information management solutions, on-premises or in the cloud. For more information about OpenText (NASDAQ: OTEX, TSX: OTEX) visit: opentext.com.

Connect with us:

- OpenText CEO Mark Barrenechea's blog
- Twitter | LinkedIn

opentext.com/contact