Retention and Disposition of Structured Data: The Next Frontier for Records Managers

Establishing a relationship with IT managers and learning about the basics of system technology will help the records manager lead the way in directing efforts to retain important data and shape policies and procedures

Laurie L. Gingrich, CRM, and Brian D. Morris, CPA

At the Core

This article

- Introduces the issues surrounding retention of structured data
- Describes the technologies IT managers use to manage structured data
- Emphasizes the importance of a collaborative working relationship with IT managers

Records managers have never questioned their need to be involved in the management of electronic records. If it was difficult to convince information technology (IT) of this, it was due to indifference rather than hostility. The first concern of IT has always been operational: ensuring system performance and availability and providing for recovery in the event of disaster. As long as the system was up and running, IT took a sledge hammer approach to records retention issues – back everything up and keep it on tape as long as needed. This allowed the proverbial killing of two birds with one stone: disaster recovery and record retention. For a time it was difficult to justify retention management of electronic records based purely on operational concerns because storage was cheaper than the labor it would have taken to design and implement retention and disposition strategies. Understandably, IT had to focus on its directive of managing systems operations.

The world has changed. There is no longer the sense that electronic storage can infinitely expand at little cost. Because of the past decade's data explosion, IT has had to routinely deal with archiving data off production systems to other storage spaces. With multiple data locales, IT faces the challenge of managing various permutations of online, nearline and offline data stores, ultimately with the question of when and how to destroy or purge them.

Operational issues are no longer the only drivers. In the landscape of regulation, risk, and compliance that global companies now face, most IT managers realize the importance of keeping records as long as needed and no longer. The risks of storing data too long on backup tapes, for instance, are well known and often publicized. Most IT professionals are now well beyond the practice of using tape as a long-term storage mechanism; they are aware that backups are part of a disaster recovery and business continuity strategy, not a retention strategy. Backing up servers to tape and storing the tapes in a closet as a retention strategy is mostly a thing of the past. Records managers can be IT's allies and colleagues in defining the regulatory and risk management justification for applying retention rules to electronic records.

Records managers are becoming increasingly adept at, and aware of, the issues surrounding management of unstructured electronic data such as desktop-created records and e-mail. They are usually less directly involved with the management of structured data, for example, records kept in accounting and other database systems. This article introduces the issues and describes the technologies IT uses to manage the retention of structured data so that records managers can collaborate with IT in developing retention policy for structured data that is consistent with the company's records management policies and procedures.

Unstructured and Structured Electronic Data

Using a classic definition, unstructured electronic records consist of electronic information created or obtained by end users where the information is not stored in tables in a relational database system. Unstructured records in a typical organization would include e-mails, word processing documents, spreadsheets, presentations and graphics – documents mostly created by individual users from desktop applications. Unstructured records would also include Adobe PDF files and electronic captures of facsimiles as well as other image files.

Therefore, by deduction, structured data is defined as data stored in fields and rows in tables of a relational database. For purposes of this discussion, the focus is on structured data that is maintained by the central IT organization. Examples would include databases containing accounting and financial data, customer data, and personnel data. (Smaller databases created and maintained by departments or individuals have their own records retention issues associated with them and will not be discussed here.)

Records managers are often involved with the management of unstructured electronic records. To cite an example that transcends industry, records managers are often involved in the selection, implementation, or management of e-mail archiving systems that work with

Glossary of Terms

Where they appear in this article, these terms refer to their IT definitions rather than the definitions typically used in records management:

Document is defined as a set of related data elements – including text, metadata, or hyperlinks – that, when taken as a whole in a business context, represents evidence of a business process or decision. All documents are considered to be records of some sort, as they provide evidence of a business process or decision and are producible in a lawsuit.

Data or data elements refer to the pieces of information that are related and collected to form the document.

Archiving is used in the IT sense of moving information off the live system into a different storage area or onto a different medium (i.e., "archiving to tape"). It is not used to mean selecting records for long-term preservation.

More Structured or Unstructured Records?

Traditionally, most company records fall into the unstructured category. According to a May 2004 Forrester Research paper cited in *Accounting Today*, 80 percent of all enterprise information exists as unstructured data.

With the introduction of enterprise resource planning (ERP) systems and the explosion of the Internet during the past 10 years, however, more business records are being found in the structured category. Estimates abound as to the percentage of structured versus unstructured data that exists within an organization. Statistics in trade publications and marketing literature vary depending on the author and the purpose of the communication.

The clear conclusion is that the volume of structured records are on the rise, but the percentage will also vary by type of organization. A financial services organization, for example, may maintain a much higher percentage of their data in structured databases, while a professional services firm such as a law firm may generate mostly unstructured documents. Most commonly, estimates are that structured data comprise 20-35 percent of the data generated in an organization.

Source: "The Next Frontier: Giving Form to Unstructured Data." ***Accounting Today*****, May 2, 2005.**

Outlook or Lotus Notes to make it easier to apply retention rules to e-mails. They are also often involved in the decision-making and implementation of document management systems originally designed to manage the storage and retention of unstructured electronic documents. Although records managers may never be as directly involved with managing records in enterprise-wide, structured databases, which is an IT function, they do need to be involved in oversight and retention issues for these systems.

Example of a Typical Record in a Structured Database System

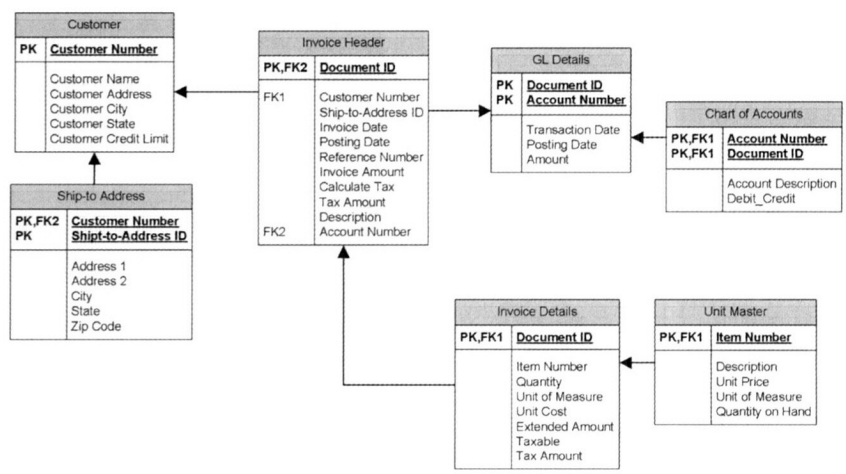
Every organization has an accounting system that can serve as an example of a structured database. Enterprise resource planning (ERP) systems combine accounting information with a broad array of other related business information, including materials management, production scheduling, and distribution, into a database of dizzying complexity. For example, the SAP ERP system, used by many large U.S. companies to manage financial data, has nearly 10,000 tables in a typical installation. Typically, these systems are managed by an underlying database management system – Oracle, Microsoft SQL Server, and IBM's DB2 are the most common.

To understand the basic structure of a relational database system and the records created by one, take the simple example of an invoice. An imaging system may contain a picture of the invoice, and that picture can be managed much like the paper copy of the invoice – stored somewhere, with a retention rule applied to it, and destroyed all at once. However, in the structured database, an electronic equivalent of the invoice also exists – but it's broken up into pieces. The customer identifier and name are stored in one table. The various customer addresses are stored in another. The date of the transaction and the total amount of the invoice are stored in another. A description of the items purchased is stored in another. All these tables are linked in various relationships. (See Figure 1)

The way the invoice appears on the screen of the end user is simply an assemblage of these various data elements. (See Figures 2 and 3)

It is up to IT to develop a workable strategy so that the data elements that the invoice comprises are archived (i.e., moved to a storage location off the production system) and ultimately disposed of at the proper time and in accordance with the organization's retention schedule for invoices. The records manager needs to know what IT is doing in regard to the invoices, as he or she is responsible for monitoring overall retention compliance. But it's harder to do in the electronic environment.

Figure 1 – An invoice as it appears in a database diagram (simplified representation)



Issues in Managing Structured Records

Issue: Defining the "Record"

Unstructured electronic records are largely electronic versions of paper records. For example, a paper contract between two parties may have its electronic counterpart in a word processing file or an image. The metadata for each may be different, but the different renditions of the record could be handled in the same manner in terms of retention. If a retention and disposition rule can be applied to a paper record, the same rule can theoretically be applied to its electronic equivalent. Problems that might be encountered in doing so are often technology-based and not related to the issue of defining the record.

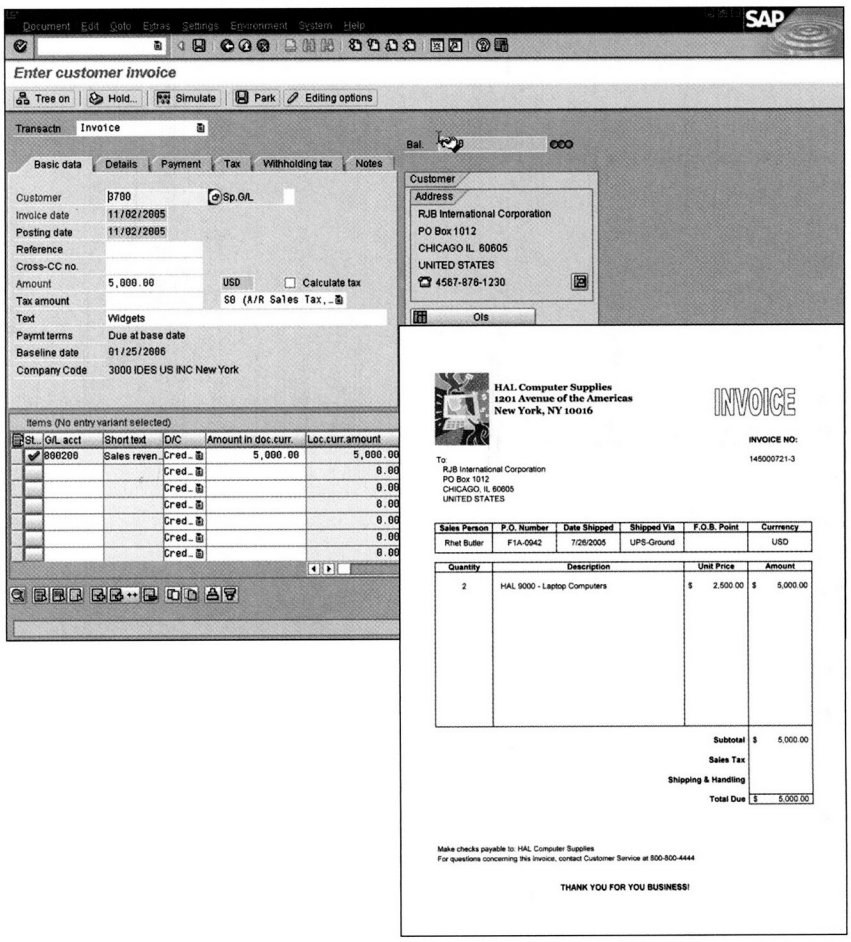
Defining the record in the structured data world, however, is precisely the intellectual problem for records managers and IT staff. Because of the way data is stored in the structured data world, the "record" or "document" does not exist independent of the technology. According to current records management theory as expressed in the international records management standard, *ISO15489-1 Information and Documentation - Records Management - Part I: General*, the characteristics of a record are structure, authenticity, reliability, integrity, and usability:

- Structure: The underlying data elements are assembled into a record of an invoice that has structure – the same data elements are presented the same way to the user for all invoices.
- Authenticity: The system will often capture information that indicates when the invoice was created, modified, and by whom.
- Reliability: The underlying data elements have not been accessed or modified by unauthorized users.
- Usability: The software/hardware combination must be able to assemble the data elements into a record suitable for viewing by humans; neither the software nor hardware can be obsolete.

Figure 2: Example of rows in a database table representing the data elements of the invoice (the customer)

CUST_NUM	CUST_NAME	CUST_ADDR1	CUST_CITY	CUST_STATE	CUST_CRED_LMT
000046532	ACME Novelty Company	1929 W. Walnut Avenue	Ace	OH	$100,000.00
000046545	Richard's Jewlry Speciality	8615 N. Port Road	Milwaukee	WI	$150,000.00
000046555	RJB International Corporation	PO Box 1012	New York	NY	$50,000.00
000046565	Nelson Tax & Associates	113 Germantown Ave	Philadelphia	PA	$75,000.00
000046568	Whett and Company	1235 W. 48th Street	Denver	CO	$250,000.00
000046569	ABC Architects	123 N Daleware Ave	Temp	PA	$25,000.00
000046570	Emma Bull	123 West 48 th street	Denver	CO	$25,000.00
000046571	Kenny A Chadburn	661 28 Street	Boulder	CO	$25,000.00
000046572	John Evans	60 Lakeview Drive	Nederland	CO	$25,000.00
000046573	Roger Zahn	12990 Central Northeast	Albuquerque	NM	$25,000.00
000046574	Laurel K. Hardin	610 8th Street	Greely	CO	$25,000.00
000046575	Chelsa Quinn Yates	4650 Pan American Freeway NE	Albuquerque	NM	$25,000.00
000046576	Andrew Williams	1030 Rio Grande Boulevard	Albuquerque	NM	$25,000.00
000046578	Mark W. Telep	30 Winrock Center	Albuquerque	NM	$25,000.00
000046579	Nancy Hagan	380 West Rainbow	Salida	CO	$25,000.00
000046580	Kevin Strauss	1600 Mechem Drive	Ruidoso	NM	$25,000.00
000046581	Charles Scott	2650 East Highway 24	Torrey	UT	$25,000.00
000046582	Patricia Kress	340 Main Street	Tonopah	NV	$25,000.00
000046583	Pat Davis	150 West South Temple	Salt Lake City	UT	$25,000.00
000046584	Mercedes Lee	1520 Fifth Street	Hawthorne	NV	$25,000.00
000046585	Robert Rankin	370 North Main	Tooele	UT	$25,000.00
000046586	Jack Norman	860 South and East	Lehi	UT	$25,000.00
000046587	Joe Masson	1525 South Street	Salina	UT	$25,000.00
000046588	Kevin Andrew Mundy	900 North Main	Fillmore	UT	$25,000.00
000046589	Tracy Collins	720 West Lionshead Circle	Vail	CO	$25,000.00
000046590	Alex Lynch	25 Vista San Pedro	Edgewood	NM	$25,000.00
000046591	Steve Martin	1217 Hudson Ave	Salt Lake City	UT	$25,000.00
000046592	Agnes Varda	1765 Dell Range Blvd	Cheyenne	CO	$25,000.00
000046593	Albert Brooks	456 Paine Street	Colorado Springs	CO	$25,000.00
000046594	Edward Burns	1402 Prospect Ave	Helena	MT	$25,000.00
000046595	Peter King	1425 Perkins Way	Seattle	WA	$25,000.00
000046596	Douglas Barker	128 Lincoln Drive	Minneapolis	MN	$25,000.00
000046597	Johny Morgan	456 Hudson Avenue	Colorado Springs	CO	$25,000.00
000046598	Marcelo da Silva	Rua Augusta 420	Sao Paolo - Conselacao	SP	$25,000.00

Figure 3: Invoice as represented to the user of the database and an imaged copy of the invoice



With structured data systems, these elements are present, but the technology provides the framework for meeting these requirements.

Issue: Proliferation of Data in Multiple Systems

In addition to the problem of defining the record in the source database, dealing with the proliferation of that record in other systems complicates matters. For example, the data elements of an invoice are stored in a structured database, and although the invoice resides in the ERP system, all or parts of it may be stored in other systems. As depicted in Figure 4, it is not uncommon for that data to be copied to a data warehouse (typically used for management decision-making and forecasting) or copied to any number of department applications or databases that use the invoices for regulatory compliance or audit defense.

In the example of the invoice, it may be of little concern that an invoice in the source system is destroyed at a certain time while a copy of that invoice remains in a data warehouse. Consider, however, a bank customer record or a patient record in a healthcare organization: the records manager may have to know in exactly which systems the records reside, as well as which data elements are proliferated in other systems, because of privacy considerations.

Issue: Risks of Non-Compliance

The risks involved in not complying with the retention schedule are different for structured and unstructured data. In the unstructured world, the management of data is decentralized among many users and there tend to be few controls. The risk of users accidentally or deliberately destroying unstructured data that should be retained is high.

In the structured environment, there are usually well-developed controls for the data's management and retention, and the risk that data will be accidentally or deliberately destroyed is lower. Data tends to be saved longer than necessary rather than be destroyed too soon, so the risks can be operational, in that too much is paid for storage. It is highly unlikely that someone in IT would purge structured data from a working database without a well-thought-out plan. The risk of losing data in structured systems mainly occurs during system migrations when IT decides not to transfer older data – sometimes without doing due diligence in investigating its continuing use and without reference to the company retention schedule.

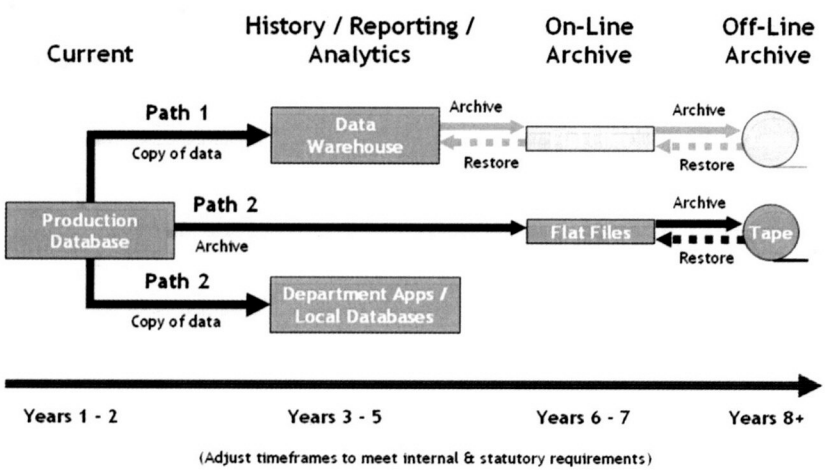
Issue: Long-term Management/ Retention

An example drawn from the real life experience of a practicing records manager illustrates the problems in managing records stored as structured data in databases. A large public school system is in the process of converting hundreds of student records from paper to electronic media. All sorts of student data elements will be stored in different fields in the same database. By law, some data – name, date of birth, address, grades - must be maintained for 60 years after the student leaves the system. Some data – special education information or honors received – can be destroyed per state law but does not necessarily have to be destroyed. Some questions facing the records manger and IT staff include:

- How can they predict the future cost of designing, testing, implementing, and monitoring a process to destroy only certain pieces of the student records versus the cost of just keeping everything?
- How can they predict the future historical value of the information? Even though it can be destroyed, should it be kept for research purposes? And, if so, should identifying information for individual students be deleted?
- How can they predict the costs of planning for future migrations for long-term preservation? A suggested alternative to endless refreshing or migrating of the data has been printing out records to paper for long-term preservation

The long-term preservation problems of this records manager would apply to any electronic records management system – structured or unstructured. Unique to the structured system, however, is having to break what was once considered as a whole "student record" into its separate component pieces of information, treating each element distinctly, and tracking where each piece of information may go.

Figure 4: Structured data "records" can exist in multiple locations across the IT landscape.



Processes and Tools for Managing Data/Records in Structured Systems

Archiving

As data volumes grow in ERP systems and databases, system performance, as measured by response time for the user, declines. To manage performance degradation, IT operations man-

Leading Practices for Data Archiving

1. Get company leadership and end-user buy-in early in the process.
2. Consider the business process and legal aspects of archiving, not just the technology viewpoint.
3. Analyze business processes to identify the interdependencies among data elements.
4. Identify closed business transactions in the database and categorize transactions so that each has predefined archiving constraints.
5. Set data retention policy that is tailored to each country. Integrate the data retention policies for each country into one archiving system.
6. Archive monthly, segregating each month's files for ease of disposition when appropriate.
7. Store archived data in classes appropriate to its retention period, access, and legal requirements.
8. Enforce data retention based on a published, central retention schedule.
9. If possible, maintain transaction transparency for users.
10. Data that does not need to be retained should be assigned a retention period of 0 years and should follow the standard archiving process.

agement may recommend moving data from the production system when the data no longer needs to be retrieved immediately. This can be done because the need to access detailed financial data – and some other types of data – declines over time. Another reason to move data from the production system may be the increasing costs of storing the data on production servers.

Moving data from the live system to an archive is analogous to the paper world's sending files to offsite storage. Archiving should be a business decision as much as a technical decision. To assess business considerations, the IT department will work directly with end users to determine the point at which records are accessed less frequently and can be archived. Records managers can be valuable partners to IT in this process by helping to define business and regulatory requirements. They can refer to the established retention schedules for paper records, for example, or assist IT with conducting an inventory of data and records.

IT and records managers are not the only stakeholders who need to be involved in this data archiving process. The company's legal counsel should be involved to help interpret laws and provide input on which records to retain, in what format, and for how long.

Some ERP systems have built-in archiving tools for long-term storage. Others require the use of third-party software tools from vendors. These vendors' solutions are not limited to ERP archiving but can also be used for archiving data from other database applications. [Editor's Note: For an explanation of how third-party data archiving tools work, see "Innovations in Information Management Technologies" in *The Information Management Journal*, January/February 2004.]

Records managers and IT management must consider how data archiving works and how compliance with retention is affected. With ERP systems, data is typically archived once the transaction is considered "business closed" and has been resident in the system for a defined period of time. If archiving takes place periodically and new archive files are created with each archive run, it is possible that two transactions that were created on the same day but closed six months apart may not end up in the same archive file. On the surface, this does not seem to be a problem. But consider the need to retrieve all detailed invoice data for a given period – as may be the case with a tax audit. With the data potentially existing in many archive files, retrieving and reporting on it can present significant challenges. It is critical that these issues be considered when designing an archiving strategy. Capabilities for future searching, retrieving, and reporting also must be part of the archiving strategy – whether it is for tax audits, regulatory compliance, or litigation.

Another concern in developing the archiving plan is the ability to destroy the records at the appropriate time with the appropriate approvals. IT performs archiving to reduce the amount of data on the production system and thereby improve system performance for end users. Eventual destruction of data is often not considered.

As a "leading practices" example for archive design, one company has taken the approach of defining the required retention period for each of its various types of structured records. When data is archived from the system, it is placed in an archive pool that corresponds to the retention period. Archiving at this particular company occurs on a monthly basis rather than the typical annual basis. Archive files are created for each month and each retention period. Therefore, each month, new archive files are created and archives expiring that month are quickly and easily destroyed. (See sidebar: "Leading Practices in Data Archiving.")

The IT Inventory

An IT inventory of systems and data is a key tool for managing structured data retention, as information gathered in the inventory will be incorporated into the

organization's retention schedule. Finding where structured data resides, however, can be challenging because, as noted, the records and corresponding renditions of the records may be managed in a number of different systems by different people in different departments. Yet, it is imperative when doing a records inventory to include structured data.

One of the main tools used for an IT inventory is still the old-fashioned interview: ask people what data they have in the systems they access, who uses it, what reports are generated, how data flows into the system, and what data flows out. Often, IT has detailed diagrams of this system information that can be used as a basis for a more detailed inventory.

Enterprise Content Management (ECM) Systems

Finally, it is worthy to note here that ECM systems are becoming much more sophisticated and comprehensive as they develop from their roots as document management systems. ECMs are increasingly designed to address structured data archiving and retention. Functionality within these systems has already been developed to apply retention rules to data in structured systems. The intent is to provide tools to purge data from the structured system at the end of its retention period with proper approvals whether that data is in the production system or archived.

To adequately oversee the retention of structured data in databases, records managers must understand the issues and be knowledgeable about the particular systems used in their organizations. In addition, they must establish a relationship with IT that promotes the establishment of procedures enabling them to obtain information needed to monitor compliance of structured data retention with the organization's overall policy and retention schedule.

***Laurie L. Gingrich, CRM**, is a manager in the Information and Document Retention practice at PricewaterhouseCoopers. She brings 18 years of experience as a records manager to a practice that focuses on records management and retention issues in Fortune 500 companies, consulting on matters of policy, procedure, and technology on the enterprise level. She can be reached at* laurie.gingrich@us.pwc.com.

***Brian D. Morris, CPA**, is a director in the Information and Document Retention practice at PricewaterhouseCoopers. With a 19-year career in information technology, he currently focuses on records management and retention of structured data for Fortune 500 companies, consulting on matters of policy, procedure and technology on the enterprise level. He can be reached at* brian.d.morris@us.pwc.com.