



X-Archive White Paper

Version 2.03

Public

October, 2008



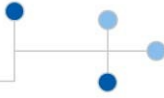
Address

Tel.

Fax

www

Itämerenkatu 5 A FIN-00180 Helsinki +358 9 5860 760 +358 9 5860 7660 www.avaintec.com



Digital Archives for Digital Records

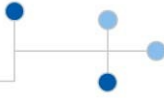
Paperless systems for managing information flows have become increasingly common. Businesses, governmental units, healthcare services, banks, and other similar organizations routinely use digital solutions in serving their customers or managing their internal information flows. These solutions vary from direct interfaces to the Internet to internal document management systems or extremely complex integrated systems that pull together data from a multitude of sources, display them to the user in a coherent, concatenated form, and even write it back to the systems behind the screen.

Such solutions notably increased organizational efficiency. Electronically stored information can be easily retrieved over networks when needed regardless of its physical location. Safeguards can be constructed to prevent unauthorized access, and the information is protected from accidental destruction by the mere existence of the system. Logistical costs are reduced, security is increased, and service quality is improved.

1.1. Why Are Digitally Originated Documents Archived on Paper?

Yet much of this information ends up printed on paper, registered, classified, and locked away in a filing cabinet in a climate-controlled vault. This entails all the logistical and labour costs involved with managing paper documents and cancels out the advantages of digitising the information flow.





So, paradoxically, while digital information flows have become the norm, paper archives are more active than ever – the increase in the quantity of information produced by the digital systems has created an overflow of paper that needs to be registered, classified, and stored in a vault, and, when needed, physically located, picked up, and copied or transported to the recipient. The logistical and fixed costs of paper archives are enormous, and yet so far there has been no adequate method for long-term archival of data in digital format.

Paper records are archived because they might be needed in the future. Very often, there are legal obligations to archive certain types of records – land records, patient data, financial information, governmental documents, etc. Not everything needs to be stored in an archive, and not everything that goes into an archive needs to stay there forever. On the contrary, a central element of good archive management is the destruction of records scheduled to be destroyed.

1.2. What Is Required of a Digital Archive?

Apart from the requirements inherent in any system built to store digital data securely, there are three specific issues to resolve in the archival of digital data. These are the *future readability* of the format in which the data is stored, the means of *verifying the integrity* of the stored data at any point in time, no matter what has happened to it in the meantime, and managing the *life cycle* of the records stored in the archive. Moreover, the legal status of digital archives has until recently been undefined: besides being technically infeasible to give up the paper archive in storing digitally originated records, it has also been legally impossible. Solutions to these problems are finally beginning to emerge.

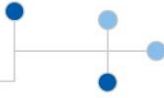
2. Resolving the Dilemma of the Archival of Digital Data

Two technological and social developments over the recent years are making it possible to resolve the dilemma of long-term archival of digital data. The first is the wide adoption of public key cryptographic technology (PKI), and the second the stabilization and acceptance of rich standards for representing digital data, in particular XML and the standards and technologies developed around it.

2.1. PKI: Non-Repudiation and Security for Data

Digital signatures are based on certificates issued by a trusted third party – a governmental institution, bank, large corporation that makes a business out of being trusted. It guarantees that the certificate has been issued to the specific individual it purports to represent. When a digital signature is calculated using this certificate and the data that needs to be signed, it is possible at any time in the future to prove that the data was signed with the stated certificate, and to verify from the certificate issuer that the certificate was a legitimate one. This means reaching a high degree of non-repudiation – in fact, a degree significantly higher than with traditional signatures, which are comparatively easy to forge and to deny.

Digital signatures do introduce one specific issue that paper signatures do not have: the expiry of certificates. Because algorithms increase in sophistication and computers in power, there will be a time in the future when any certificate



issued today will be cracked – that is, it will become possible to forge signatures identical to the authentic ones. The solution to this problem is to limit the validity period of digital certificates. They are typically issued for two to five years, after which they must be replaced by new ones, the new certificates being complex enough to resist cracking for their period of validity. If we simply store signed data in a database, fifty years from now there will be no way of telling whether or not the information has been tampered with.

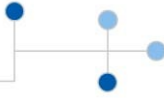
The solution to this problem is to issue a certificate to the information system that stores the data. As the certificates of the signatures in the system expire, the system will re-sign the data using its own certificate, which is periodically replaced. This way, we can be certain that all the data in the system has been signed with a valid, secure certificate. By maintaining this unbroken chain of signatures, we can guarantee the integrity of the data indefinitely.

2.2. XML: Making Data Semantic

The second technological and social development has been to *put data first*, i.e. the introduction of rich, standard formats for representing digital data. In traditional information systems, logic comes first, and data is something for the logic to operate on. This has resulted in the proliferation of mutually incompatible data formats – information that is only readable if the program with which it was created is available. From the long-term storage point of view, this is a real problem: as the programs and the systems they run on become obsolete, the data itself will become unreadable, no matter how carefully it is stored.

This problem is finding its solution in the emergence of international standards for representing digital data: starting from encodings (such as Unicode, a unified scheme for representing any character in any language in use on the planet) to higher-level abstractions. In particular, XML has spawned a large number of more or less rigorous and useful standards with which data can be represented semantically: with the meaning either contained within the data itself, or in a *schema*, a rigorous machine- and human-readable description of the data. With some thought and consideration, almost any form of structured data can be represented as XML in such a way that it is understandable regardless of the data system used to store or read it. If these methods are used scrupulously and intelligently when setting up a digital archive – that is, the representations of the archived data are semantic, defined as machine- and human-readable schemas, and all of the data that will be archived is stored in the accepted formats – the problem of data obsolescence can be solved. Even if the format itself becomes obsolete, it will always be easy to convert it to whatever output format is needed because its semantics are explicitly defined.

Finally, the legal framework for adopting digital archives for digital data has been enhanced by new standards for records management. Now organisations are better able to understand what the technical and legal requirements for a digital archive are, and utilise tools that support these standards.



2.3. The Digital Records Schedule: Automating Record Lifecycle Management

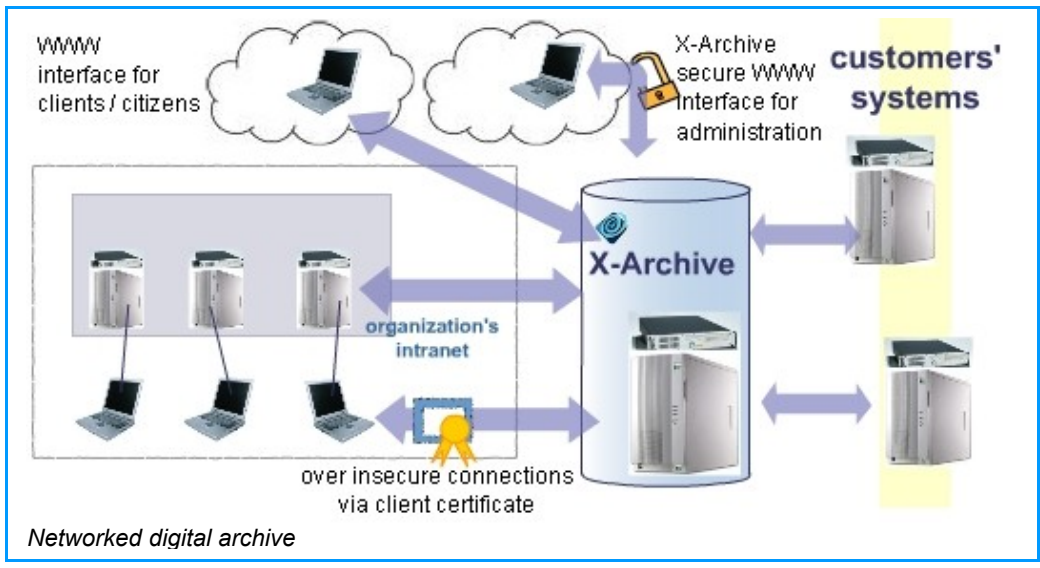
The *records schedule* is a traditional tool for archive management. It contains information about archived records, and rules about how they should be treated: who is allowed access to them, when they should be destroyed, which records take priority in case of an emergency such as a fire or a flood. Often these rules change during a record's archival period. For example, a record may originally be confidential, but may become public after a certain amount of time has passed since its creation. These rules are often legally binding.

In a paper archive, one of the main duties of the archivist is to interpret these rules: when a record is requested, it is his or her job to evaluate whether the individual requesting the record should be permitted access to it, to see to it that records get destroyed as the rules demand, and so on. Unfortunately, the increasing volume of archived records and the sometimes undue complexity of the rules set down in the records schedule have often resulted in situations where especially rules about timely destruction are observed patchily or not at all: this does not only contravene laws, but is also highly inefficient, as storage of these unnecessary records contributes to the running costs of the archive.

If the records schedule is digitized in such a way that these instructions are expressed in a form that is unambiguous, clear, and both machine- and human-readable, it becomes possible to automate the management of the record life cycle. State changes, such as a confidential record becoming public, can be computed automatically from record metadata, such as the creation date, archival date, or data subject birth. Destruction lists can be compiled by evaluating the instructions against the same record metadata, so they only need to be examined and approved by the archivist, greatly easing his or her workload. These efficiency gains apply just as well to physical as well as digital records.

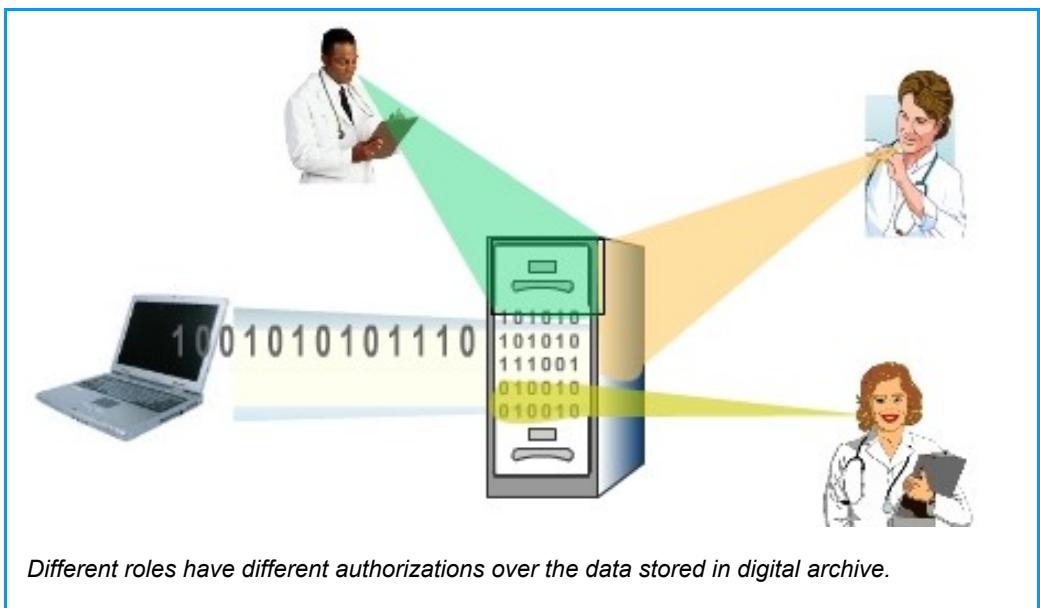
3. A Comprehensive, Customizable, Networked Digital Archive

To fully realise the productivity potential inherent in paperless systems, the data produced must also be archived in digital format. The systems that create the records must be able to send them into a digital archive for long-term storage. This way, organizations can gradually phase out costly, cumbersome, and heavy paper archives, which results in significantly improved productivity, savings in logistical and running costs, and easier accessibility and retrieval of required information. Citizens, patients, and customers can access the records related to them, while having these records safe from unauthorized access by others. The avoidance of labour and other costs associated with the management of a paper archive will quickly pay back the investment in a digital archive system. Digital data needs to go into a digital archive, and X-Archive is a comprehensive, customizable solution for a networked digital archive. It is undergoing a process for certification as a Records Management system under ISO standards.

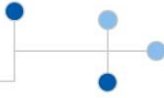


3.1. Administering X-Archive

X-Archive is a comprehensive solution for a digital archive, based on these technologies and approaches. It includes a full-featured archive management tool, which is based on a WWW interface and requires no client software. Its digital records schedule automates the normal functions associated with archive management, such as registration of archived records, access control and access logging, classification with a structured system of records schedules, storage and retrieval, state changes during a record's archival period, and scheduled destruction.



Different roles have different authorizations over the data stored in digital archive.

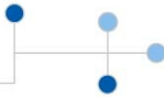


A record may only exist in a single “real” or “primary” classification at a time, since the rules governing the record's life cycle must be unambiguous. Therefore, X-Archive permits the creation of *secondary classifications*. A secondary classification is an alternative classification tree that contains pointers to classes in one or more virtual archives. This makes it possible for a record to exist virtually in more than one classification scheme. An organization's virtual archives may evolve along with the organization, but external actors such as information systems or client organizations may be access subsets of the archives through a secondary classification that stays stable. This means that should the real classification system change, it will not be necessary to make corresponding changes to integration tools or practices used by client systems or organizations: it is enough to change the mappings centrally in the secondary classifications they use.

For example, suppose two municipalities decide to merge a part of their functions. Formerly, each of them has stored the records created by that function in their separate virtual archives. After the organizational change, they will be using a new, shared virtual archive. Now, if each municipality's client systems and organizations have accessed the virtual archives through a secondary classification, all that needs to be done is to change the mappings in these secondary classifications so that new records end up in the new virtual archive, while searches encompass both the old and the new classifications.

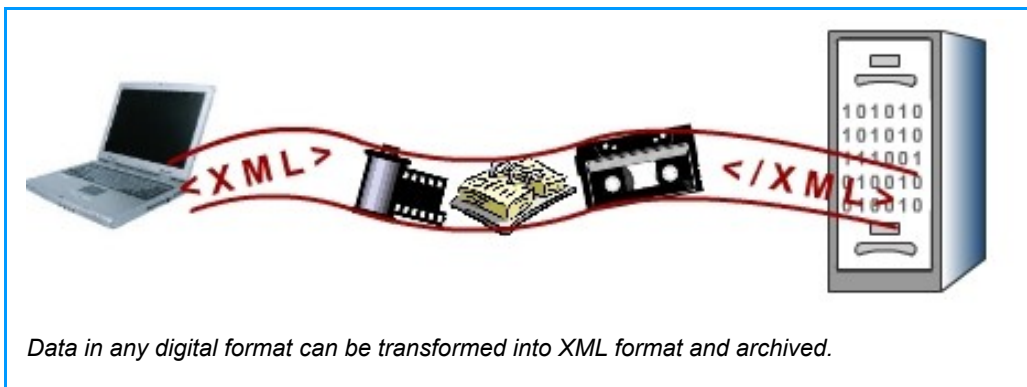
Access control and authorization in X-Archive is based on a policy model of user categories, data categories, actions, reasons, and obligations. This permits flexible administration of privacy policy within an organization – access to individual records or even parts of individual records can be defined by defining data categories and mapping them to user categories and data subjects.

All archive requests are logged to a set of registers, so that a permanent record of archive activities is created: it is possible to audit every event at any time in the future. Digital signatures can be required from each archive request – typically of committal of incoming records, destruction orders, and other operations affecting the contents of the archive. Checkouts of sensitive records can be required to be signed as well. The system maintains the integrity of the archived data by periodically re-signing it with the system certificate.



3.2. What Digital Data Is Archivable?

X-Archive is highly flexible. In principle, any digital data can be archived. The metadata structure is defined as an XML schema and is fully extensible and customizable; any vocabulary, such as Dublin Core, can be used to express it.



Individual records are identified with a globally unique identifier, for example the OID system, although any global identification scheme can be used.

X-Archive is designed to scale to fully accommodate even very large digital records, up to several gigabytes each, even for computationally intensive tasks such as computing hashes for signatures. It is therefore suitable even for archiving digital media files such as audio, video, pictures, or laboratory data.

The classification scheme and data descriptions are fully customisable: it is one of the primary functions of the archive manager to expand on these schemes and descriptions, as new types of data need to be archived. The archived XML data can also be associated with XSL style sheets and other resources used for displaying it. It is possible to define several separate “virtual archives” running on the same physical X-Archive server. These appear to the archivists and users as completely independent systems. This makes X-Archive highly suited for use as a centralized archive solution, e.g. in regional or local government, hospital districts, or large corporations.

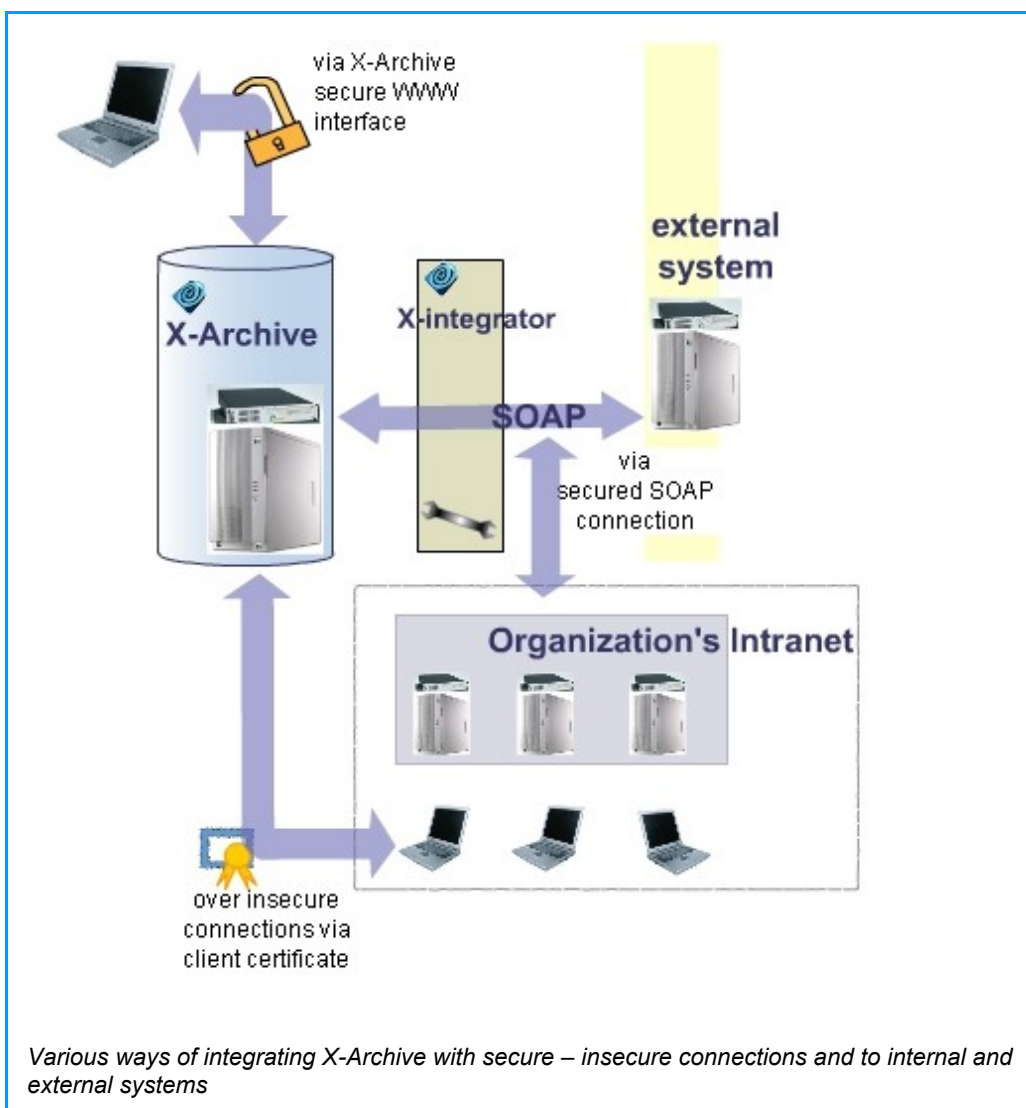
3.3. X-Archive in Existing Information Environments

X-Archive can be accessed both using its own secure WWW interface, or directly by other information systems over a secured HTTP connection. By securing the connections with client certificate authentication and other means, it is possible to create a trusted machine web over existing, insecure connections, even the Internet. X-Archive is designed to function equally well in a geographically distributed context as locally. Additionally, integration tools can be used to wrap existing information systems so that they can communicate with X-Archive using its HTTP protocol.



X-Archive is designed to scale to handle large volumes of records and requests. It incorporates a request queuing system that permits operation in asynchronous mode during periods of peak load, or to handle lower-priority requests such as periodic batch committals that would otherwise crowd out higher-priority traffic. It is therefore well suited for environments such as hospitals, with a large number of information systems producing and requesting large volumes of records and peak periods of increased load.

If the archived data is in XML format it can be associated with XSLT stylesheets for different presentation formats. This makes it easy to use X-Archive in a multichannel environment: the client application simply retrieves a record from X-Archive, and the associated style sheet is used to transform the record into, for example, WML for display in a PDA, HTML for display on a web browser, or PDF for printing.





X-Archive can also be used as a repository for records in active use rather than for long-term storage, if transparent handling of digitally signed records and fine-grained, high-security access control in real time is needed.

4. Modular System Based on Open Standards

X-Archive is a modular system based on industry-standard platforms. The server components are written entirely in Java on a component framework; it can run on any platform that implements Java 1.4. The database connectors are modularised as data access objects that contain SQL queries tailored for individual database systems. This means that X-Archive can be adapted to function on any SQL-based relational database management system. X-Archive is designed for easy installation on clustered or load-balanced systems; its different components (HTTP front-end, archive management client, archive core, database system) can be collocated or distributed across several machines. It is easy to tailor a digital archive solution to fit the load, usage, access security, and data protection level required for the application under consideration, from single-server systems to clusters or mainframe installations serving large volumes of data in multiple virtual archives.

The X-Archive administrative client is XML-based and fully customisable. It generates an HTML/ECMAScript interface. It provides easy access to the archive's administrative functions. The daily operation of the archive is very simple: an archivist needs no more than a few hours of training to be able to administer a digital archive using X-Archive because the conceptual basis is familiar.

5. Standards and platforms

- **Administration Utility:**
 - Operating Systems supported and tested
 - Windows 2000
 - Windows XP
 - Linux
 - Microsoft Internet Explorer, version 5.0 and later
 - Firefox, version 1.5 and later
- **Server:**
 - Operating Systems
 - Linux (kernel 2.4 and 2.6)
 - Windows NT, version 4.0 and later
 - Windows 2000
 - Windows XP
 - Windows Server 2003
 - Java 2 SDK 1.4.2 (included as part of installation)
 - Jetty 5.1.6 (Included as part of installation)
 - Apache Cocoon 2.1.4 (Included as part of installation)
- **Tested databases:**
 - Microsoft SQL Server 2000
 - PostgreSQL, versions 7.4 to 8.1

- 
- Oracle database, version 10g

Supported standards:

- XML 1.0
- W3C XML Digital Signature (XMLDSIG) standard
- XML Schema
- JDBC 2.0
- Java 2
- ANSI SQL
- HTML 4.01
- JavaScript 1.2
- CSS Level 1
- X.509
- CDA Level 1 and Level 2
- HL7

6. About Avain Technologies

Avain Technologies develops and tailors user-friendly, networked, secure and flexibly managed ready-made products for companies and organisations, allowing them to implement their service and process flow management and store their information in digital form in the long term. Our solutions enable a phased transition to a fully digital, paperless environment.

Our technology is based on open industry standards, and we participate actively in international standardization forums.

By working in close and interactive cooperation with our customers and partners, we develop products that allow an increasing number of companies and organisations to improve their operations and services, reduce costs and increase the commitment and contentment of both personnel and customers.

Our customers include organizations in local and national government, healthcare, and finance sectors.

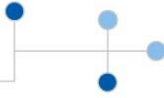
6.1. Avaintec's values

Our key values are:

- ensuring customer satisfaction,
- respecting the individual in everything and everywhere, and
- offering advanced technology with high information security.

6.2. Ownership

Avain Technologies Ltd is owned by its employees and institutional investors, such as The Finnish National Fund for Research and Development (Sitra) <http://www.sitra.fi/eng/index.asp>.



6.3. Date of Foundation and Company History

Avain Technologies is a Finnish software company founded in 1997. Its main products are X-Archive, an XML and PKI-based solution for long-term archival of electronic records, X-Web Form Manager, an XML based solution for serving, submitting, signing, and processing electronic forms securely over the Internet, and X-Digital Signature Suite, a component that can be integrated in WWW-based software providing digital signature functionality using PKI, GSM, and other technologies.

6.4. Contact Information

Company:

Avain Technologies Inc.
Itämerenkatu 5 A
FI-00180 Helsinki
Finland

Tel: +358 9 5860 760

Fax +358 9 5860 7660

WWW: <http://www.avaintec.com/>

E-mail: info@avaintec.com