ArchiSafe DLM Forum Budapest, Paper (Proposal)

ArchiSafe a legally secure and scalable long-term record keeping strategy

The impact of modern digital information processing environments has been remarkably universal, extending to industry, government, and the academy; to business, science, engineering and public administration as well. The Personal Computer has conquered the public offices also for a long time already. Paper bound files and the traditional office as an essential organizational basic unit of classic administration loses more and more meaning in view of the full coverage introduction of IT supported processes. E-Government, yesterday still in-word, is already the perfect example of a counter model to the classic paper based bureaucracy. Vast quantities of information in digital form are produced, exchanged, and used in a variety of settings, for myriad purposes. All of these diverse applications of digital information technology rest on a common foundation of shared benefits, including powerful search and retrieval capabilities, network delivery, perfect duplication, and interoperability.

At the same time, all efforts to electronically upgrade the public administration will lose the inherent benefit when leaving the gap between the front and back office, i.e. if furthermore standing files, fax and file cards prevail in the so-called back office in the background. The back office is the central production place of the administration, the organization of the "working state". The central product of the back office is the file. So the construction of modern administration structures on an electronic basis will also concern the file primarily. The now electronic file shall support in the future the unhindered and fast exchange of data and information between the administration units to the advantage of the citizens and the economy particularly.

But, just as the benefits of digital information environments transcend people, systems, and domains, so do the challenges which accompany them. Nowhere is this more evident than in regard to preservation of digital information, i. e. securing the long-term persistence, integrity and authenticity of information in digital form. The essential advantage of digital information to be machine-readably immediately when digitized and coded and in consequence as simple bits and bytes may be transported even about large distances within a few seconds only is also the decisive weakness. Digital information due to their virtuality isn't only fugitive but can be manipulated also very easy and unnoticed. So anyone who promotes the electronic file as an essential precondition for the further progress in the electronic government has to deal with the challenges to guarantee the authenticity and integrity of electronic documents.

As a central principle of the public administration any essential communication has to be fixed in files so that everybody at any time will be able to reconstruct the administrational decisions. As a rule, this principle finds his expression in the respective business and official regulations by the regulation that "the state of an administrational matter must be recognizable completely from the files" at any time. This central commandment of the administration behavior also applies undoubtedly to the electronic file. Within the framework of E-Government particularly electronic documents and electronic records need legally secure and enduring status for long time, despite the fact that computing systems (hardware and software as well) have short live and electronic information may be easily corrupted. And exactly this is the aim of the project ArchiSafe.

Thus, based on international (e.g. www.prov.vic.gov.au/vers) and national (www.archisig.de) experiences with electronic record keeping strategies, the Physikalisch-Technische Bundesanstalt (PTB), a federal research institute under the auspices of the German Federal Ministry of Economics and Labor, has started in 2005 a project for developing and implementing a legally secure and scalable electronic archival solution which not only grants the enduring status of electronic documents for long time but also complies with the requirements of the German electronic signature law.

The scheduled solution is indicated by the following features: As a first focal point, instead of taking a system oriented approach to electronic documents and files, a more data oriented approach seems more appropriate. Thus, electronic documents generated by some IT-application and dedicated to long-term preservation will be converted into specified and standardized data formats (PDF, TIFF) and encapsulated in a layered (onion) XML structure, which provides support for layers of metadata (onion model) of process information. The layers itself are defined by XML schemata, which not only enable the archival system to examine the correctness of the internal structure of the XML data packages to be archived but also the user to add some metadata specific to its own business domain and purposes. The data structure is adopted to encapsulate the document (payload), the processing context, and authentication in a single object.

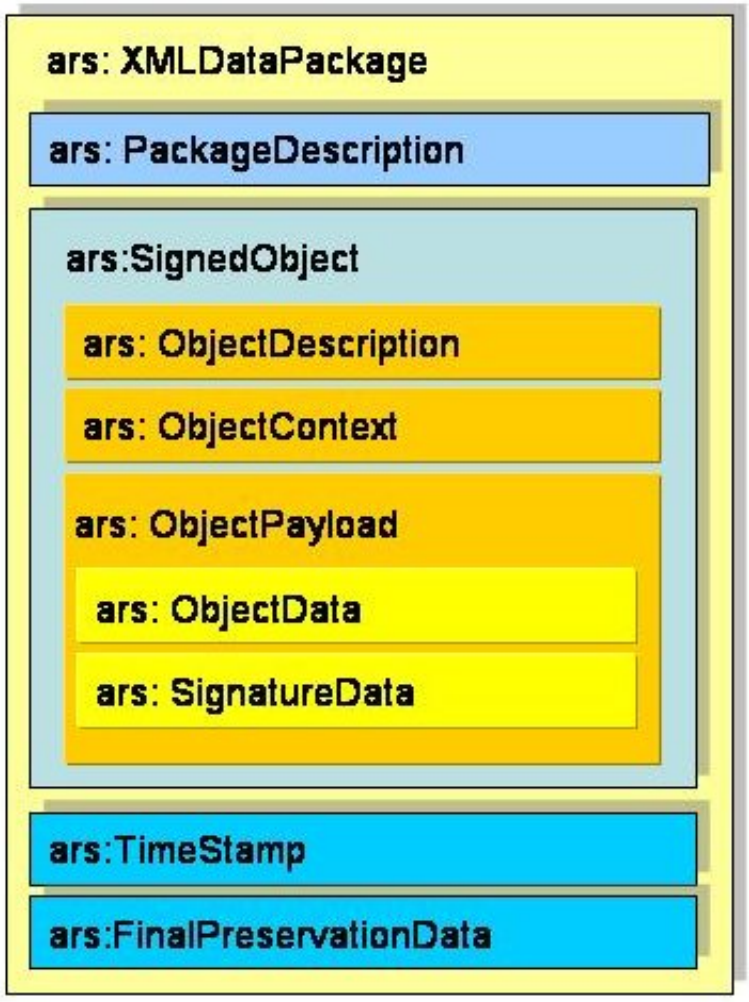

Fig. 1: ArchiSafe principal data structure

The underlying design of the data structure matches with the following principles:

- Self-documenting: it should be possible to interpret and understand the information in the data package, at least at a primitive level, without reference to any external documentation.
- Self-containing: the structure should contain all information about the content and the process. It is far easier and more reliable to manage the information if it stored in one place rather than in components which are stored separately. Therefore documents will be Base64 encoded for embedding it in the xml structure, while database tables can be encoded directly using xml.
- Extensible: it should be easy to extend or adopt the structure for adding new metadata or metadata containers without affecting the interoperability of the basic design.

The primary selection criteria for representing document content using PDF and TIFF format was the confidence that, for the foreseeable future, it would be possible to write a viewer for the document from publicly available information (longevity of format support). Besides this, by 2006 an ISO version of the PDF-A format will be available, a standard which specifies a limited, vendor-independent subset of PDF for documents that is especially designed for long-term archives. The use of TIFF for physical document capture for document archiving will continue to be a safe choice. In addition, embedding a TIFF image into a PDF-A format will provide greater business value in many cases, such when complex searches or comprehensive audit trails are needed.

The use of XML provides an open and standard way of recording the structure of documents XML is a standard derived from an ISO technology standard designed in the mid-1980s called Standardized General Markup Language (SGML). This ISO standard, which originated from document structuring technologies developed at IBM in the early 1970s, is a mature approach for structuring documents. The XML document format provides the capability to use an independently defined XML vocabulary for representing the structure of digital documents. By 2008, major application vendors like IBM and Microsoft will introduce XML-based document formats that can be easily transformed into any archive format.

Thus, the combination of a vendor-independent archive format like PDF-A and a more vendor-independent document format based on XML provides a maturing approach to the need for a more application and platform independent document format. In addition, these future-oriented formats will allow to more easily complying with regulatory needs regarding electronic records preservation.

The second focal point is that electronic documents dedicated to long-term preservation should be captured and protected at the time of creation or time of receipt. Using of electronic signatures and time stamps grants for the integrity, authenticity and non-repudiation of the electronic documents and will be archived as well as the evaluation results of the signature data (validity of certificates at time of evaluation). This way the legal validity of the stored documents is safeguarded durably.

Inspite of this, you have to keep in mind, that electronic signatures are based on cryptographic algorithms, keys and certificates. Opposed to hard copies cryptographic algorithms and keys lose their safety suitability in the course of the time and electronic certificates will be documented from the corresponding certificate authorities for time periods no longer than 30 years. ArchiSafe therefore will capture and store any data qualifying the electronic signatures and signature certificates just at the time of archiving and renewal the signatures periodically.

Figure 2 shows the principal system architecture of the ArchiSafe environment. The archive request is initiated by any domain application (e.g. a document management system). The application generates the schema conformed xml data package, and fills the tags with all information needed to describe and maintain the content of the document just as the context of the document (process data). After that the xml data package will be processed by the ArchiSafe middleware, which calls a cryptoprovider for evaluating signature data and affixing an initial time stamp as a cryptographic seal to the whole package. In the following the xml package is send to the archive system which generates an archive unique identifier (AUID) as a "cloakroom ticket" for uniquely identifying the document in the storage system and the domain application as well. The AUID therefore will be given back to the domain application.

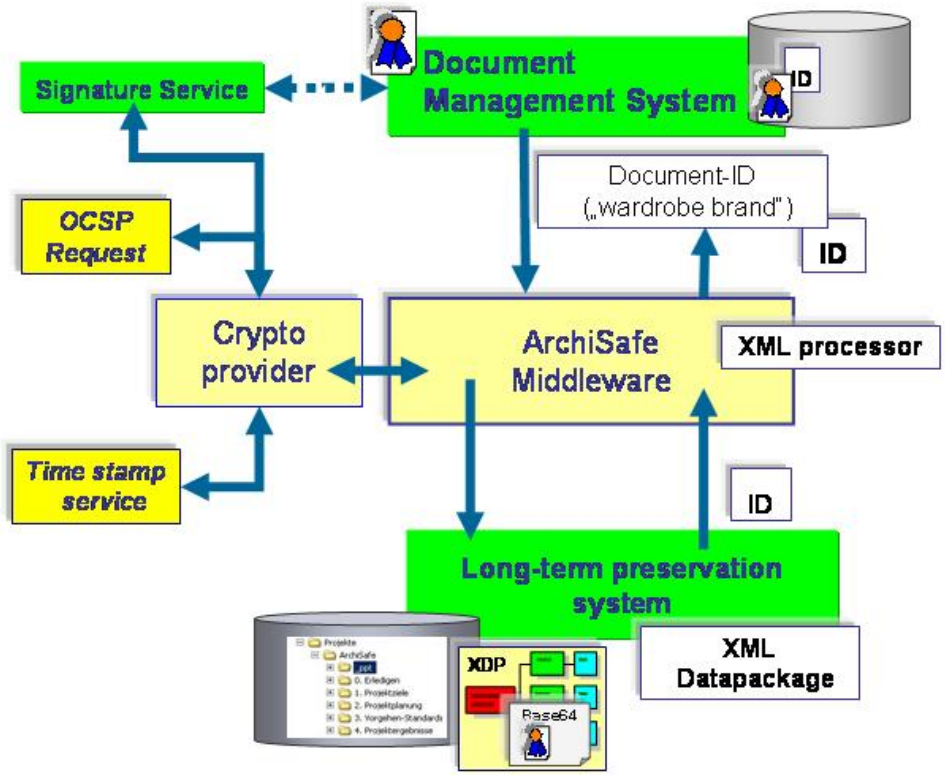

Fig. 2: the Principal ArchiSafe Architecture

To maintain the legally secure status of the archived data over a long time and for achieving a high-performance and economic archive solution with respect to the regularly required renewal of the signatures especially, the scheduled archive system builds hash-trees from a configurable amount of archived data packages (www.archisig.de), which subsequently will be time-stamped on the top of the tree only. Thus, the renewal of signatures, owed to a possible break of the underlying cryptographic algorithms, is only required for the time stamp at the top of the tree. The use of hash-trees in addition offers the opportunity to remove single documents from the storage, e.g. for reasons of the protection of data privacy. As long as the hash-tree remains intact, the remaining documents in the storage do not leave their legal validity.

How does such a tree work? The application which causes the electronic documents and signatures saves the document including the signature and all accompanying information in the

archive system. When putting down the archive system calculates the hash values of every file, like the digital "fingerprint" of the document. From these individual hash values now a hash-tree is built up, and the top is “sealed” with a time stamp. Such hash values have primarily two qualities important to this task.

- A hash is unique, i.e. every change in the file also leads to inevitably changes of the original calculated hash value. Manipulations of the file therefore don’t remain undiscovered.
- The calculation of a hash is a one way function, i.e. one cannot infer from the knowledge of the hash on the contents of the file. This is primarily important if for legal reasons single documents must be deleted. One can delete single documents straight away from the archives with that. As long as the accompanying hash remains in the archives, the seal is still valid over all "branches" and "leaves" of the hash-tree but you cannot infer from the existence of the hash in the tree on the contents of the deleted document.

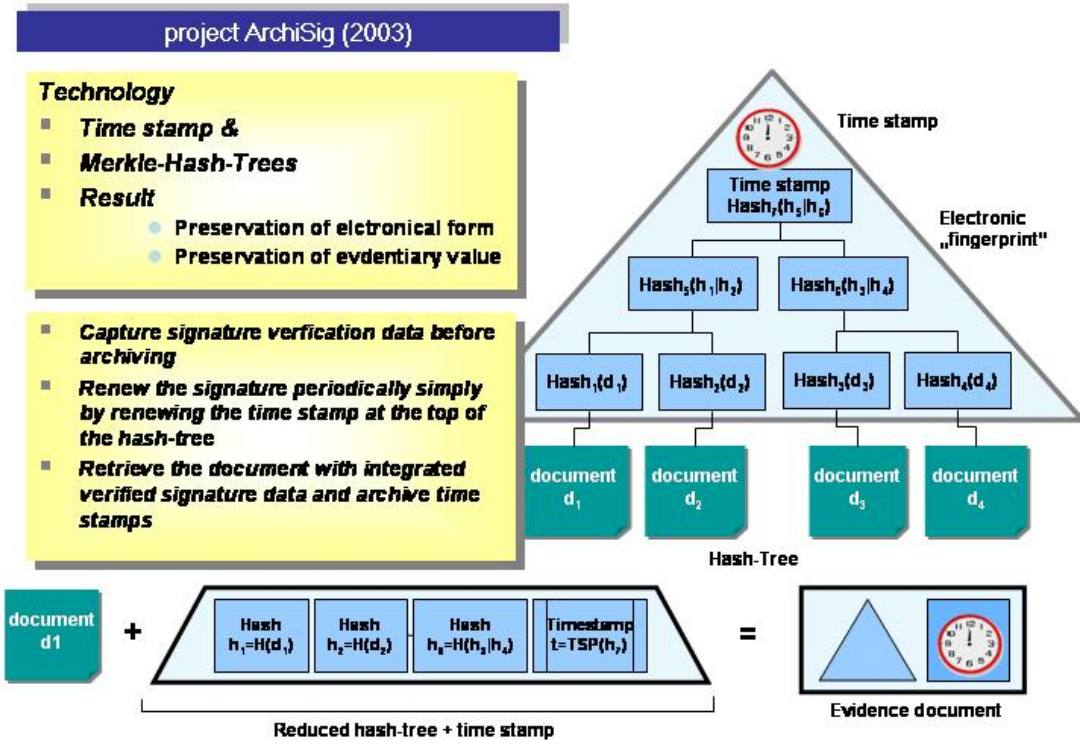

Figure 3: Merkle hash-trees for long-term preservation

What is won with that now, however, because the hash and "seals" are for their part also based on the use of "aging" algorithms? The signature renewal and the safety risk were reduced to an easily comprehensible number from "seals" of a variety of documents which lock the individual hash-trees. As long as these "tree seals" aren't broken, the individual hash values and thus also the documents are safe.

Any time a signature renewal must be carried out now, this means that merely the "seals" must be replaced. The hash values in the "trees" aren't concerned by it. Any unnoticed manipulation of the single documents and signatures can be excluded furthermore. The risk of a complete hash-tree renewal still can in addition be reduced by it that two redundant trees are

built up with different algorithms. If an algorithm gets unsafe now, the second tree is still safe and it remains plenty of time so to calculate the hash values of all documents in the first tree newly.

The project ArchiSafe is supported and promoted by the E-Government-Initiative “BundOnline 2005” of the German Federal Government and published on www.archisafe.de.