FROM IT-SYSTEMS TO THE ARCHIVES

Jan Dalsten Sørensen *

UDC: 004.3:930.253

***Jan Dalsten Sørensen: From IT-systems to the archives**. Technical and Field Related Problems of Traditional and Electronic Archiving. Conference Proceedings, Maribor 4/2005, No. 1, pp. 109-114.*

Original in English, abstract in English, summary in German.

This article discusses the challenges presented by the digitization of government and it describes how the Danish National Archives handles the appraisal and transfer of e-records. It also points out some of the areas of development and focus in the years to come.[1]

INTRODUCTION

Everywhere, the development toward more and more e-government presents the archives with a major challenge: When more and more important information is found in digital form, how can we make sure that this digital information is identified, transferred and preserved in a way so that it will be accessible and meaningful in the future? What changes in archival practice are necessary in order to meet the demands of this new situation? How do we handle problems like technological obsolescence? How do we go from just recognizing the challenges and to actively deal with the e-records in a way that will preserve the information that is worthy of preservation for future generations?

The Danish National Archives received the first transfer of e-records in the mid-1970s, has a long tradition of making demands on the structure and use of document managing systems of the state authorities and was in 1995 the first archives to draw up specifications for the transfer of electronic document management systems, so that the authorities that wanted it could dispose of their paper records.

IDENTIFICATION, APPRAISAL AND APPROVAL

We have found it useful to divide the IT-systems that are used by the authorities into the three following basic categories.

- Databases. These can contain almost any kind of information gathered by the authorities in connection with their business, but they are not used for document management purposes.

* ***Jan Dalsten Sørensen, Special consultant, Head of Section for Electronic Records, Department of Appraisal and Transfer Address: Rigsarkivet, Rigsdagsgården 9, DK-1218 København K, Denmark.***

[1] *The article is a revised and extended version of my paper "Public Regulations for Electronic Records: Regulations and Requirements for Public Authorities in Denmark" that was presented at the International Congress on Archives, Vienna 2004.*

- Traditional document management systems (DMS) that digitally store metadata about documents and case files but where the actual documents are physical paper documents and thus not part of the electronic system
- Electronic document management systems (EDMS) where both the documents themselves and the relevant metadata are stored digitally.

We define electronic records, or e-records, in the broadest possible sense, namely as digital information stored in a system of one of these above-mentioned categories. The underlying understanding of the Danish archival legislation is that "records" is to be understood as all the information that any authority gathers and stores in a structured way while performing its duties.

When it comes to paper records, we as archivists usually do not see them until they are 20 to 30 years old, or maybe even older. And that is fine. Obviously, it is necessary that the authorities have procedures that ensure that records are made and kept in an orderly manner, both logically and physically. But usually, we do not have to worry about the paper records as long as they are still in active use or the first 10 or 20 years after that.

With e-records, it is a different story. It would be a huge challenge to try to capture data from IT-systems that are 20 to 30 years old, migrate them to current standards and transfer them to the archives for long-term preservation. There is a huge risk of technological obsolescence, for instance in terms of obsolete formats or the loss of data due to the use of hardware that is no longer supported, or any number of other situations that might put the data at risk.

It is also necessary to take the European data protection directive into consideration. The provisions of the data protection directive – at least as implemented in Danish legislation - means that personal data cannot be stored in the systems of the authorities any longer than necessary. Often, this means that data must be deleted no later than 5 years after the last contact with the person in question. However, this information is often of historical importance, which is recognized by the Danish data protection act. The solution is that such information must be transferred to an archive before deletion. This compels the archives to receive early transfers of data – and in order to do that, we need to identify those records as early as possible. Therefore, when a Danish state authority takes a new system into use, it must notify the National Archives.

If the new system is a database without any document management functions, it is sufficient for the authority to notify the National Archives two weeks before the database is taken into use. The notification must be made using a special form and include information about the purpose of the system as well as a description of the data in the system, along with other information that is necessary for the appraisal process.

Databases are not subject to approval by the National Archives but the notification takes place so that the databases can be appraised as early as possible. Provided that the data is found worthy of preservation, the date for the first delivery of data from the system concerned is set at the time of appraisal. Usually, the first transfer takes place after 5 years of operation. We inform the authority in question about the technical specifications for the transfer, so that they have ample time to figure out if there are any problems concerning the conversion of data from their particular database into an archival version.

If the system is a document management system, it has to go through a process not only of appraisal but also of approval. Therefore, the notification on new document management systems has to take place at least three months before the system is taken into use.

The notification must be made using a standard form and include

- a description of the retrieval system (e.g. filing plan, index terms, metadata required to retrieve documents)
- draft instructions for the use of the system
- technical documentation

If a system is implemented as anything but a relational database, the authority must already at the time of approval give us a description of how the system can be converted to a relational database at least in the first normal form because that is how it is eventually transferred to us.

The system must be divided into archival periods of approximately 5 years. When the archival period expires, all case files must be closed, and the database with metadata and, if applicable, digital documents, must be closed and an archival copy must be transferred to the National Archives. If the case files are on paper, all files from the period must be set aside as a coherent series, and new files must be created for the new archival period.

The application of the system must be well defined, e.g. whether the system is used by all of the authority or only certain parts thereof. If parts of the organization use the system as a traditional document management system where the case files are paper based, while other parts of the organization use it as a fully electronic document management system with all documents in digital form, it is vital that this is well defined, too.

Even if the fully digital document management system is used universally within the organization it is necessary that it is possible to register documents that are found on paper (e.g. odd size documents that cannot be scanned into the system or reports where you only scan the front and the table of contents). Therefore, all systems must have a field where you can register whether the document is stored fully digitally, partially digitally, or on paper.

The system must be implemented in a way that ensures the preservation of the documents that are stored in the system. This means for instance, that we need to know exactly what document formats can be stored (Word, PDF, TIFF, Excel spreadsheet etc.). Only documents that can be converted to TIFF must be found in the system and if we are in doubt we demand that the authority present us with a statement from the software contractor that only documents that the contractor will guarantee the conversion of can be stored. We do recommend that the conversion to TIFF takes place as soon after the registration of the document as possible. All of this must be included in the instructions for use so that all users know what document formats to store in system, when they are converted to TIFF etc.

We also demand that the authority takes other precautions against the loss or corruption of information and these precautions must also be described in the instructions. A good example is the annotations that you can use in Word documents: At the conversion to TIFF, the information would either be lost or the annotation

would hide some of the text in the document. A possible solution is to store the information as endnotes when the document is converted to TIFF.

Another example could be that we require that the authority describes how they will ensure the documents if applications like their word processing software changes during the archival period. Here, a good solution would be to convert all documents to TIFF before the changing of the software.

The retrieval system must be described in the instructions as well and it should be made to ensure that the documents that belong together are found together. The instructions should describe the metadata that must be found for documents and files.

It should be noted that it is not possible for a software contractor to get an approval of a particular document management system as such. We approve of the implementation of a system, and we consider the technical solution together with instructions, filing plan etc. as a whole. A well-working system could be implemented in a very inexpedient manner – and thus it is not enough only to look at the system itself.

The requirement for approval of new systems only applies to state authorities. However, also local governments are obliged under the Archives Act to ensure that e-records are preserved in a way so that they can be transferred to a public archive. As of May 2004, the technical specifications and requirements for the transfer of data to the National Archives apply to the transfer of all public data. That is regardless of whether the data comes from at state or a local authority and regardless of which archive the data is actually transferred to.

TRANSFER

When data is transferred to the National Archives it must be migrated to a system independent format based strictly on standards. Some of the main points are,

- Data must be transferred as a system independent archival version
- data must be structured as a relational database
- there must be sufficient metadata to describe structure and content
- general information about the system must be included
- the documents must be converted to a standardized format
- when the system in question is a document management system, some SQL-queries must be included
- and finally, the data must be transferred on CD-R or a portable USB-hard disk.

Data from each table must be converted to this standard character set (ISO 8859 – Latin 1) and stored as sequential files.

The consequence of migration to a system independent archival version is the loss of the functionality of the software application that created data. Therefore it is important to transfer data from the systems with sufficient metadata and context information. Among the most important parts of the metadata is a machine readable XML-file with a description of all tables and all fields, based on EBNF (Extended Bachus-Naur Form). Thus, we make sure that the tables of any archival version, no

matter which system generated them, are described in exactly the same way. This is, of course, necessary in order to be able to make some sort of a standardized retrieval tool for all the archived data.

A metadata description of the tables is not enough to guarantee the future reuse of the data. It is necessary to ensure additional documentation and a selection of general information must therefore be included in the archival version. It has not been possible to make an actual standard for this since we receive data from all kinds of systems where the types of documentation available and necessary are equally diverse. However, we try to make sure that the general information includes an administrative description of the system as well as technical documentation, which illustrates the structure and functionality of the system, e.g. ER-diagrams and screen dumps. All this information must be stored as TIFF-files.

Specific archival information is also found in the form of descriptive tables with data about e. g. provenance, the period of time in which the data was generated and the name of the system. There is also a table that gives an index of the documents providing general information.

As mentioned before, all documents must be converted to TIFF. Documents must be compressed.

If the system includes sound- or video files, they must be archived as MP3 for audio and MPEG-2 for video.

Finally, if the system is a document management system, 3 to 10 of the most used queries must be included in the metadata in order to help future user find the right information when the original application is, of course, no longer available.

When data is transferred, it is critical that it is tested so that we can be sure that the all the requirements are met. The authority cannot delete data in its own IT-systems until the National Archives has completed the test and issued a letter of approval of the transfer, and thus taken the responsibility for the future preservation of data. When the archival version is OK, we burn two sets of CD-R that are stored on two different locations. An additional copy of the archival version is stored on a disk RAID (Redundant Array of Independent Disks).

AREAS OF DEVELOPMENT AND FOCUS

Seen from an archivist's point of view, the migration of data from the original system in which it was created to a system independent archival version is not an ideal solution. With regards to paper based documents and case files, we as archivists want to make sure that the information is kept and preserved authentically, i. e. the way it was made and preferably also in the same order, cf. the principle of provenance. The migration of data from the original system to a system independent archival version involves a transformation that might put the authenticity of the records in jeopardy. However, we also need to be realistic and accept that while migration to a system independent format has its downsides, at least it is a solution – and probably the only solution - that works for now.

As long as we have no real, working solution for an emulation strategy that allows us to preserve the attributes of the systems themselves, we always need to improve and refine the general information that is included in the archival versions to make up for the deficiencies of the migration to system independent archiving.

Another area of focus right now is the development of a standardized retrieval tool so that data in the archival versions can become more accessible. It is of course possible to recreate the tables in, say, an SQL-server but you cannot do that without certain qualifications. Obviously, I doubt that you will ever be able to use data from archival versions without at least some qualifications – just as you will never be able to read a handwritten document from the 17th Century without some qualifications, as well as a lot of practice. But with that said, there is a growing need for a tool that makes it a lot easier for the potential user to recreate the databases and make queries without too much programming. We plan to develop such a tool over the coming years.

ZUSAMMENFASSUNG

VON INFORMATIONSTECHNOLOGIESYSTEMEN ZU ARCHIVEN

Den dänischen Staatsarchiven wurden seit der Mitte der 70er-Jahre elektronische Akten übergeben. Gestützt auf die Erfahrungen dieser Archive betont dieses Referat die Notwendigkeit von:

- Frühere Erkennung und Bewertung elektronischer Akten als jene von Schriftgut
- Aufforderung an Behörden, für die entsprechende Anlage elektronischer Akten zu sorgen, ehe diese ans Archiv übergeben werden
- Übergabe elektronischer Akten an das Archiv zu einem frühen Zeitpunkt
- Übertragung der Daten auf ein unabhängiges archivtaugliches System, welches auf Standardformaten beruht
- Fortführung der Entwicklung von Standards, Verfahren und Hilfsmitteln, um die Qualität und den Zugang der ans Archiv übergebenen und bearbeiteten elektronischen Akten zu verbessern.

Jan Dalsten Sørensen *holds an MA in History and Latin from the University of Aarhus. Since 1999 he has worked in the Department of Appraisal and Transfer, the Section for Electronic Records, at the Danish National Archives. Since 2001 he has been the head of the section. He represents the Danish National Archives in the DLM Forum and has written several articles and presentations on the appraisal and transfer of e-records.*