Transfers and Preservation of E-archives at the National Archives of Sweden

Magnus Geber, the National Archives of Sweden

Summary

The National Archives of Sweden have received transfers of digital records since the 1970s. Today we get digital files from a lot of different systems and we also have a large amount of scanned images from our paper records to take care of. To handle this situation we have created our electronic archive system, RADAR, primarily for transfer control and preservation. This is a solution that consists of a number of modules developed at different phases. It is based on international standards, like the OAIS-model and metadata standards, as well as experiences of many years of work with digital records.

Background

The National Archives of Sweden (Riksarkivet, in short **RA**) have received transfers of digital records ever since the 1970s. Around 1980 we were among the national archives in the world, maybe the one, with highest volume of digital records. However this was handled in a very basic way, reel-to-reel tapes and paper documentation. During the first decade we didn't have any of our own equipment, we used service firms and the transfer control was made from data printouts and paper metadata. When we acquired technical tools they were basic and the work in many ways was still manual. More or less from the beginning we used a general migration strategy implemented with transfer demands concerning technical formats and documentation (metadata). RA has the right to issue regulations for the governmental agencies concerning archival matters. The preservation method included copying and conversion when required.

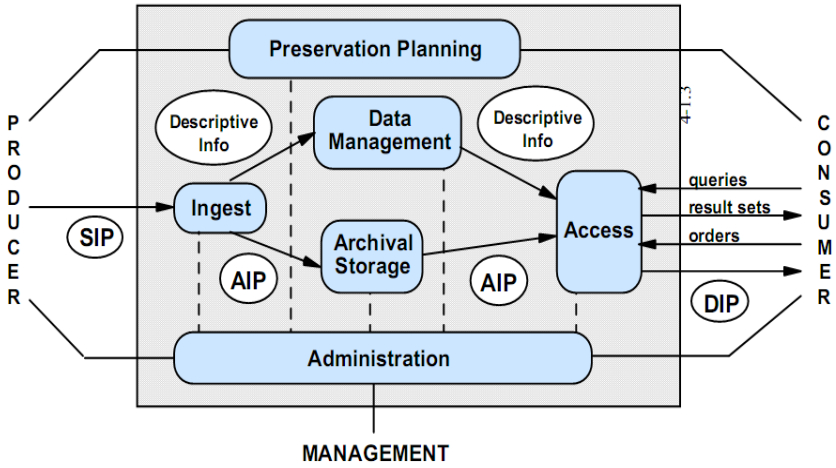
During the last decade we have been able to use more resources to develop our work with electronic archives. At the same time there has been international progress in the field. The prime example of this is the OAIS-model. But there are also a number of metadata standards as well as findings from many development projects. All this has been important input to our development process. But also our own experiences from many years of work in the field have contributed in addition to what we have learnt from our international networks, especially the one between the national archives of the Nordic Countries.

All this has produced the result that today we have a solution for the electronic archives of RA, a platform for digital preservation, consisting of a number of modules together forming a full system. The different modules have been developed at different occasions and been introduced gradually. The whole solution, called RADAR, has not yet been fully completed although many parts already being used, some parts are developed but not introduced and some still needs to be developed. But having worked with many parts of it, we can say that it will be a complete operating solution. Due to technical evolution and other external factors, we expect that further development of RADAR will continue into the future.

Prior knowledge

To read this text certain knowledge of electronic archives and digital preservation would be of use. Basic concepts are taken from the OAIS-model.[1] IP stand for Information Package and SIP, AIP and DIP means Submission IP, Archival IP and Dissemination IP. An IP is a package normally containing the data to be archived together with all relevant metadata.

Figure 1 The OAIS function model



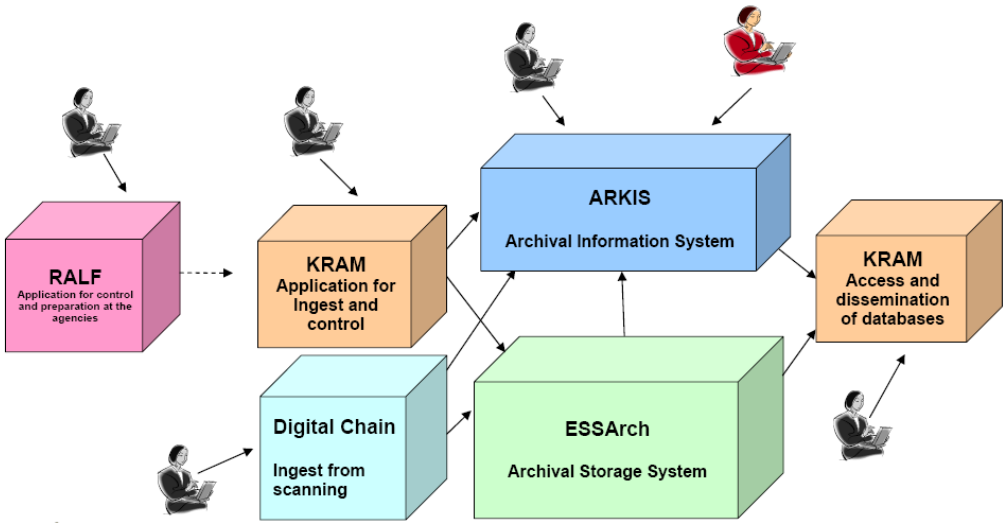
RADAR-platform

The RADAR solution for the electronic archives of RA, mainly consist of the follow modules and specifications.

- A general package model (SIP, AIP) originating from OAIS
- Use of general metadata standards such as METS and PREMIS
- Use of specific metadata standards for different types of systems, for example the ADDML-standard for databases
- A self-developed system for transfer control and conversion, KRAM
- A free downloadable self-testing tool of SIPs: RALF
- An open source storage system ESSArch
- Full integration with our general Archival Information System, ARKIS
- Work processes for different type of systems or record types, the “Digital Chain” and other

The different modules are shown in figure 2, below. These modules as well as the specification for the information packages, which not are shown in the figure, will be described in the following sections. Figure 2 is in many ways a parallel to the general OAIS function model. RALF is a tool to help agencies create a correct SIP for transfer to RA before the Ingest process. KRAM is an application consisting of tools to be used by RA in the Ingest process as well in the Preservation and Dissemination processes. The Digital Chain is the methodology for our own scanning activities resulting in digital files which will also end up in our Electronic Archive. ARKIS is the Archival Information System describing all our records, paper as well as electronic ones. ESSArch is the storage system creating and managing the digital AIPs and their physical storage.

Figure 2 **The Swedish National Archives platform for digital preservation (RADAR) 2010-12-15**

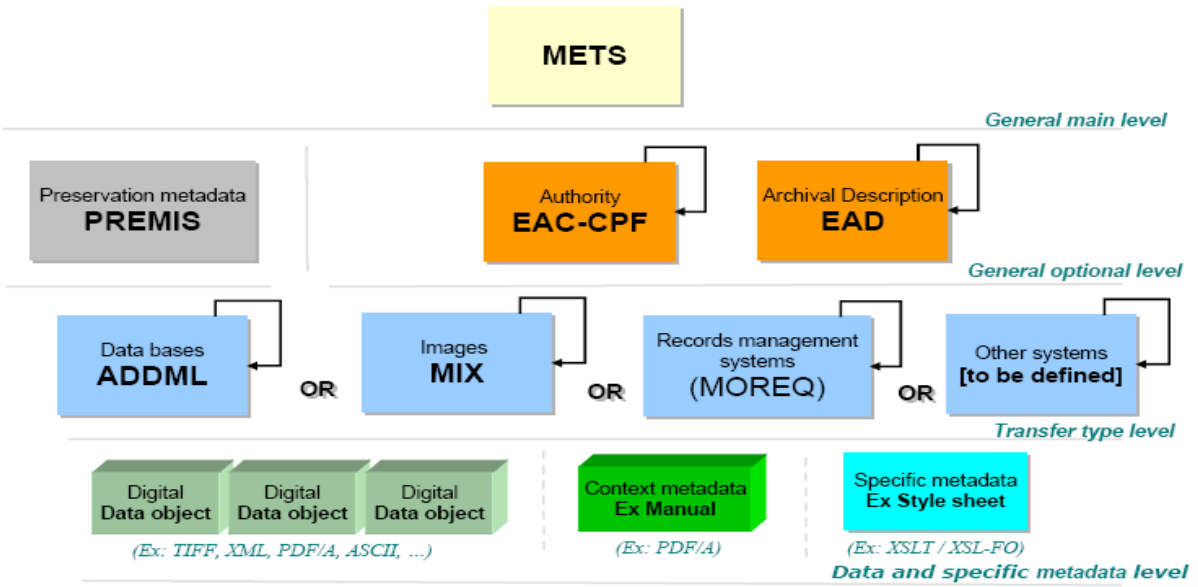


RADAR is created to be able to manage the different types of electronic files that RA is handling including our old transfers of database files, today's transfers from many types of systems including sound and video-files and the image files scanned from our paper records. We are also in the process of making RADAR compliant with the TRAC-demands (Trusted Digital Repositories and Audit Checklist) and ISO 16363:2012 [2]. TRAC formulates 90 criteria for the archive to reach the full status of trustworthiness. These technical demands involve RADAR and we fulfil parts of it already. In TRAC there are also administrative demands affecting RA in other ways.

A general package model (SIP, AIP) and General metadata standards such as METS and PREMIS

The package specifications are basically following OAIS. Our SIP- and AIP-structures are quite similar. The main idea is a general and common package structure which has different options concerning the content. The content options are dependent on the source system and we call them transfer types, which each have its individual specification (see figure 3).

Figure 3 Architecture for information packages SIP/AIP



In accordance with the OAIS model the package consist of different kind of metadata. The metadata we use is normally in XML, specified by schemas, often based on international standards. The general and common package level is basically built up by the metadata specified by the METS standard[3] describing the general information about the package. At the general level there is also metadata from the PREMIS standard[4] concerning Ingest and Preservation actions. Today this is an option for SIPs but always in our AIPs. Added to that in the general level you can also have archival descriptions concerning the content of the package according to the internationally developed EAC-CPF[5] and EAD[6] standards. Today this is an option for SIPs and not part of our AIPs. But that metadata can be imported to and exported from our Archival Information System, ARKIS, which is described below.

In the next level of the model you can see the optional metadata dependent on transfer type.

At the bottom there are the actual data files together with specific metadata files like manuals in PDF or style sheets.

Specific metadata standards for different types of systems, for example the ADDML-standard for databases

Depending on the originating system of the transfer there are different specifications. We have chosen to call these transfer types because the distinction of the types is not fully consistent but based on the practical situation of today. When there are systems designated to work with a special field like Accounting, HR or Medical Records these will be considered transfer types and there will be specifications standardizing parts of the content. But there are also other situations such as a lot of business systems in more specific fields where the system use a database. In these cases there are transfer type "databases" specifying how to describe the structure and the content but not standardizing the content. Web sites might also be an example of such technically based transfer type.
We have specified a transfer type for digital images based on the MIX-standard, mainly used for packages with our own scanned images. In a project together with the public sector, eARD, see below, we are developing transfer types for Records Management and HR-system. Others will follow.

But the transfer type which we by tradition have been handling mostly is databases. Until recently databases have formed the vast majority of our transfers from governmental agencies. If there is a database based business system in a field which not yet has got a content-based specification it can use this transfer type.

The schema we use for database transfers is called ADDML. The specification was originally developed by the National Archives of Norway together with tools for Ingest. Today there is an agreement between the national archives in our two countries concerning the further development of the schema and others are welcomed to join. What we have developed together is the common base for the standard and since the two countries have slightly different practice, we use the same standard but not the same solutions.
The schema consists of three part; reference, flatFiles and dataObject. Reference contains general metadata about the transfer and origination system. flatFiles describes the structure, technical options and content of the data base files. The database files are supposed to be in the format of flat text file. Partly this metadata can be seen as a record description of the content. DataObject is an option to append files of other formats, often as reference.

We are aware of the Swiss and PLANETS-supported SIARD-format. We believe our format is more primitive but also simpler, for example data are in flat text files and not in XML, and

may be easier to use. There might be a future development with more harmonisation between the formats and for example tools for conversions between the two.

A self-developed application for transfer control and conversion, KRAM

KRAM, is an application able to control, document and convert the transferred information packages. It can also upload archived data base files into a SQL-data base system to facilitate dissemination. The original and one of the main functions of KRAM is the control of the packages at Ingest. The metadata of the SIP is checked for completeness and then the actual data is checked for agreement with the metadata. More thorough control of file formats, using DROID or JHOVE, for example, is in our plan for the future.

The other main function of KRAM is conversion. That might take place at Ingest if the content of the SIP not is in accordance with the formats we use for our AIPs. It can also be part of our future Preservation actions when the content of old AIPs is in formats that are becoming outdated. However at present we have big backlog of outdated file formats from the 1970s and 1980s which we are converting as part of a special project, using KRAM. Today the conversion actions are implemented for the database transfer type, which are the type of all our old transfers.

After a conversion a new ADDML-file is automatically created to contain the new relevant technical metadata. All important actions done at Ingest and Preservation is also recorded in a PREMIS-file and added to the AIP.

It should also be noted that we always preserve the first generation of an IP with data files in original format.

A free downloadable self-testing tool of SIPs: RALF

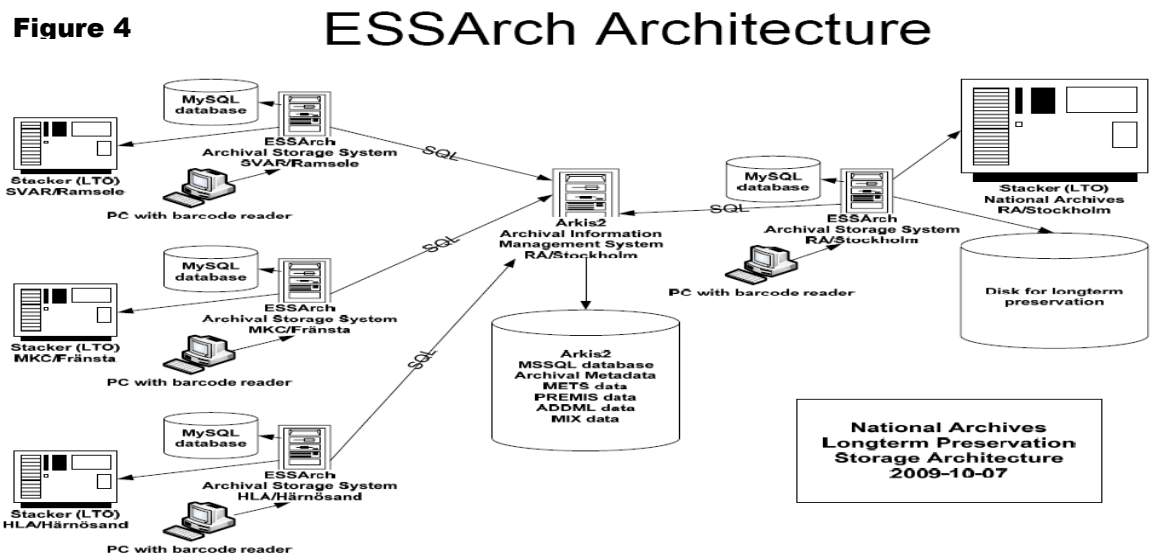
RALF is a software which essentially is a downsized version of KRAM. Its purpose is to help the governmental agencies to fulfil the transfer demands of RA before Ingest, when transferring databases. The metadata produced by the agency is checked by the software to be correct. There is also a basic check of the data files to be in accordance with the technical demands and corresponding to the metadata. For the purpose not having to manually produce an ADMML file there is an Excel-template where all the relevant metadata can be registered. RALF then converts the Excel-file to an ADDML XML-file. RALF also creates a basic SIP with METS metadata.

Today RALF runs under Windows and is programmed in .NET. It is downloadable with an installation module.

An open source storage system, ESSArch

EESArch is the Storage module in our solution. It manages the packaging, storage and retrieval of the AIPs. Checksums are used both for the packages and the contained metadata and data files. There is logging of the events taking place. The packages are in TAR-format and may be stored on different media, normally disk and tape, at present LTO. The tapes might be of-line, which mainly concerns the large TIFF-files of our scanned records where compressed copies in other format are used for dissemination. ESSArch handles multipliable copies; normally there are at least 3 copies stored at two different geographical locations. The system is distributed and decentralized to many of the localisations of RA. For examples there are subsystems at the units where the scanning takes place to be able to create AIPs in direct connection with the production. The subsystems have local MySQL-databases using the PREMIS-data model. A major part of the metadata of the local systems is exported to the central Archival Information System of RA, ARKIS.

Figure 4



ESSArch is developed by a small consultant firm in cooperation with RA. It consists of different modules and is basically Open Source. The National Archives of Norway also uses ESSArch but their implementation has a lager capacity and partly do tasks KRAM do by us.

Full integration with our general Archival Information System, ARKIS

The Archival Information System of RA, ARKIS, is managing all our records regardless if it be paper, digital etc. The part of ARKIS handling our digital records is an integrated part of our Electronic Archive, the Data Management part according to the OAIS model. The normal archival description is contained in ARKIS. There are import and export functions for EAC-CPF and EAD. The data model of ARKIS is generic and based on what we call archival units.

Archival units can be fonds, series, subseries, documents, electronic files etc. Connected to the archival unit, there are a number of attributes like physical storage, topography, chronology, accession etc. Between the separate instances of the archival unit there are different types of relations, of which some are strict hierarchal. Depending on type of archival unit there might be different metadata attached. Also a big part of the technical metadata is handled by the system. Not all technical metadata is imported to the ARKIS database; some of it is in the relevant XML-files which are stored on a server. With ARKIS these XML-files can be presented by XSL-transforming using designated style sheets.

Work processes for different type of systems or record types

There are different regulations concerning the public records produced by the governmental agencies, some issued by RA and some other addressing digital records. Many of these can be seen as regulating work processes.

Another work process concerns the material produced by RA that also ends up in RADAR. That is the scanned images of different paper records. The images are created mainly at the "scanning factories" in Fränsta and Ramsele and partly at other units within RA. After initially having had a rather chaotic situation today there is the so called "Digital Chain" process to manage the scanned production. There are demands saying that the paper records have to be registered in ARKIS before scanning takes place; that there is a plan for storage of the electronic files; and that there is a budget covering the main costs. During scanning some basic metadata concerning the archival description of the original records is put into the digital headers of the TIFF-files.

Transfers

Transfers to RA are facultative for the governmental agencies. But if they chose to keep their archival records they have to take the full responsibility for the dissemination and preservation of the records. The Swedish law concerning Freedom of Access stipulates that a vast part of the records is open for the public. The preservation issue means that an agency might have to acquire their own electronic archives solution. To avoid that many agencies choose to transfer their digital records to RA, often by the time a new business system replaces an older one.

Traditionally the transfers of digital records have been a time consuming activity. The formats of the metadata and our tools were relatively primitive and the transferred material often contained a number of errors or shortcomings. For the agencies the transfers of digital records used to be rather odd occurrences. There were also limited man recourses on both sides. To

improve the situation RA has set improved demands for machine readable metadata and developed tools to create and control SIPs. A major step has been the introduction of the excel-template for database transfers and the distributed downloadable tool RALF. In the near future we will work with full SIPs in the Ingest process. Traditionally we have set formal demands for the transfer formats. In the future we will apply a more flexible attitude using the price list for transfers, basically meaning that a complete SIP will be charged a lower price. This may also be the relevant concerning the file formats of the transferred data files, different price depending on by us preferred format. We also plan to develop new tools to support the creation of SIP by export from the original business systems. Bigger agencies having internal system containing the relevant metadata will probably prefer to create full SIPs in automated processes. The agency Statistics Sweden (SCB) is already now preparing such routine in cooperation with us. Smaller agencies will probably use our tools and if necessary pay us extra to complete the SIPs.

Today all transfers are done by data media, tape, disk, USB etc. We are also working on accepting transfers by net, for which purpose access security is the main issue to be solved.

eARD-project

Since last year RA has been leading a project together with other agencies to create common specification for electronic archives within the public sector. The project is supported by the governmental commission, the "E-delegation", and is run in cooperation with the co-organisation of the Swedish county councils and municipalities. It should here be mentioned that the county councils and municipalities exercise a degree of self-governance. They have their own liability for their archives and are never supposed to make transfers to RA. All governmental agencies will eventually make transfers to RA, but especially the big agencies will prefer to create their own electronic archives to be able to take care of their own digital records for a certain time.

The eARD project have given priority to make specification for the SIPs, using a similar model as the one used by RA; a common package model containing different kinds of transfer types with separate specifications. It derives from a national forum for institutions having e-archives, including RA. In the eARD-project some functional demands will also be specified. Based on these specifications the agencies can develop there their own applications but common application can also be created by agencies together or by contractors as a service. Having a common specification the future transfer to RA from governmental agencies with their own electronic archives will be much easier to execute. And the general exchange of information in systems all over the public sector will be facilitated in many ways.

Contributors

Many persons have been involved in the development of our solution and I want to mention a few. Mats Berggren, our main system architect, is behind many parts of the model and has also contributed with figures to this paper. Karin Bredberg is our XML and metadata expert and is behind most of the XML schemas and many other specifications in the model. Göran Kristiansson has served as project leader or manager during many of the phases ever since the first ARKIS-project. The ElArk unit, doing the practical work with the electronic archives within RA, has also made important contributions.

References

[1] ISO (2003) Open Archival Information System (OAIS), ISO 14721:2003. (2012-03-29). http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] TRAC and ISO 16363, Standard for Trusted Digital Repositories http://www.crl.edu/Archiving%20%2526%20Preservation/Digital%20Archives/Metrics%20for%20Assessing%20and%20Certifying-0

[3] Library of Congress (2010) Metadata Encoding & Transmission Standard (METS). (2012-03-29). http://www.loc.gov/standards/mets/

[4] Library of Congress (2012) Preservation Metadata: Implementation Stategies (PREMIS). (2012-05-30) http://www.loc.gov/standards/premis/

[5] EAC-CPF (Encoded Archival Context - Corporate bodies, Persons, Families) http://eac.staatsbibliothek-berlin.de/

[6] EAD (Encoded Archival Description) http://www.loc.gov/ead/