One Digital Repository to Preserve Life, the Universe and Everything – The New Electronic Archive of the Saxon State Archives

Karsten Huth; Saxon State Archives; Dresden, Saxony/Germany

Abstract

This article is a short status report concerning the new Electronic Archive of the Saxon State Archives. It contains a description of the implemented OAIS functional entities "ingest", "preservation planning", "data management" and "access". The article also mentions the importance of the organizational structure of an electronic archive.

The Beginning

Increased efforts in e-Government in the German federal state of Saxony are creating new challenges for the Saxon State Archives, which must find a way to preserve digital data as reliably and legally trustworthy as analog records.

The items produced by e-government systems are very diverse in terms of their content, structure and behavior. Normally, IT-projects have precise definitions of incoming and outgoing data, but in the field of archiving e-government data, you are forced to design your technical archival infrastructure without exact knowledge of the size, structure and format of the data you will have to preserve. Electronic records, geographic data, statistics, audio and video formats are only examples of possible items that have to be preserved within the next years and there is only one technical infrastructure for all of them. Designing the technical infrastructure under those uncertain circumstances is, of course, an almost impossible task. Organizational changes of appraisal und access in the digital worlds have to be considered as well.

The Saxon government initiated the project LeA (**L**ong-term storage and **e**lectronic **a**rchiving) in 2008. The goal of the project was to build a long-term storage system for governmental agencies to preserve e-governmental data for as long as the data is needed for administrative purposes and to build one electronic archive for data that has to be preserved forever as prescribed by the Saxon Archives Act. The project was a joint effort of by the Saxon State Archives, the Saxon State Ministry of Justice and European Affairs and the Staatsbetrieb Sächsische Informatik Dienste.

The Saxon State Archives

The Saxon State Archives is tasked by the Saxon Archives Act with preserving the records of the courts, authorities and other public bodies of the Free State of Saxony. The archives are organized into five departments at five different locations (Dresden, Wermsdorf, Leipzig, Chemnitz, Freiberg). One hundred employees serve about three thousand visitors each year. All locations together hold 103 kilometers of paper records, maps, photographs, moving pictures, sound recordings, pictures, posters and videos.

How to plan an Electronic Archive?

What kind of items should our archives hold in the future? How many items should we expect each year? How much storage do we need in the first year and what is the exact format of the data? These are essential questions for every new information system. Unfortunately, catching these important benchmark numbers before you try to set up an electronic archive is nearly impossible.

The State of Saxony has over 400 agencies, and every agency can produce data with archival value. At the beginning of the project LeA, the State Archives started a survey to collect at least some data to estimate cornerstone requirements for an electronic archive. At the end of the survey, the project team wasn't much smarter, but least we had some idea of the future. Thus we've based our concepts on the following two premises:

- OAIS [1]-based information and process models
- Organizational concepts

OAIS-based Information and Process Models

The first cornerstone of our concept is, of course, the OAIS. All electronic archives or archive systems in Germany with long-term preservation features claim to be OAIS-conform. German standards on long-term preservation, DIN 31644 [2], DIN 31645 [3] and DIN 31646 [4] are also built upon the foundation of OAIS.

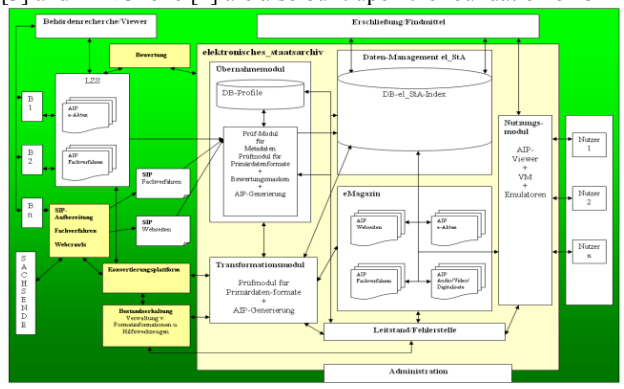

***Figure 1** Customized OAIS Functional Model*

Therefore, OAIS-terminology and the functional model are adopted in our concept, and used to describe all necessary processes und modules. Because the OAIS is a reference model, it is applicable to any kind of archives, but the functional and the information model has to be customized again and again. The functional model must be adapted to fit into the respective technical infrastructure.

For example, the functional entity archival storage, which provides services and functions for the storage maintenance and retrieval of AIPs, has to meet the requirements of our data host. Because the Staatsbetrieb Sächsische Informatik Dienste (SID) has a strategy to standardize the storage infrastructure, we had no other choice than to adapt the very open description of the functional entity in the OAIS to the standardization efforts of the SID.

Ingest

About eighty percent of our concept explains the complex functional entity "ingest". It is very important for an electronic archive to get the highest quality AIPs it can possibly get. The quality of an AIP depends very much upon the quality of the incoming SIPs. Unfortunately, the quality of the SIPs is often very bad. That's because many of the e-government systems never considered possible data transfers to an electronic archive as an important requirement. Existing e-government systems have no interfaces to automate the building of SIPs. Under these circumstances the best way for the archives to get acceptable packages with metadata and data objects is to build the SIPs on their own. Therefore, the concept describes the requirement of a pre-ingest SIP-building tool and the requirement of additional staff.

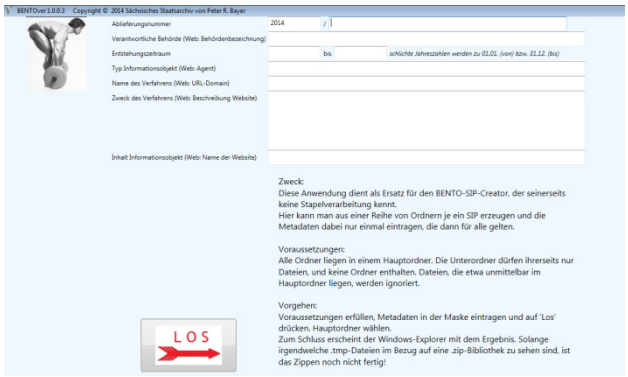

Figure 2, *Screenshot of the SIP building tool "BENTover"*

Our concept foresees three different types of ingests. At this moment, the most important type is used for data stemming from any type of e-government system without a standard interface for transferring data to an electronic archive. In these cases, the SIP building tool "BENTover" structures the exported data into SIPs. The SIPs fit the BENTO-standard (eCH-0160) invented in Switzerland. After this pre-ingest procedure, some parameters of the following ingest have to be adjusted. We have to inform the archive system which of the possible SIP standards it has to expect. The system expects only certain data formats, and maintains a list of standard scenarios and what to do in each case. So the archives staff, while setting up this list of "ingest agreements", controls which data formats in the SIPs are acceptable and which not.

The next important step is the definition of possible migrations during ingest. It's our goal to reduce the diversity of data formats in the AIPs to a reasonable minimum. Therefore, our archival system has converters for basic archival formats like PDF/A and TIFF.

After setting up the ingest agreement, the following ingest steps will proceed automatically. The system uploads the SIP and checks the integrity of the whole package. If the package isn't of the expected standard, the ingest procedure will stop immediately.

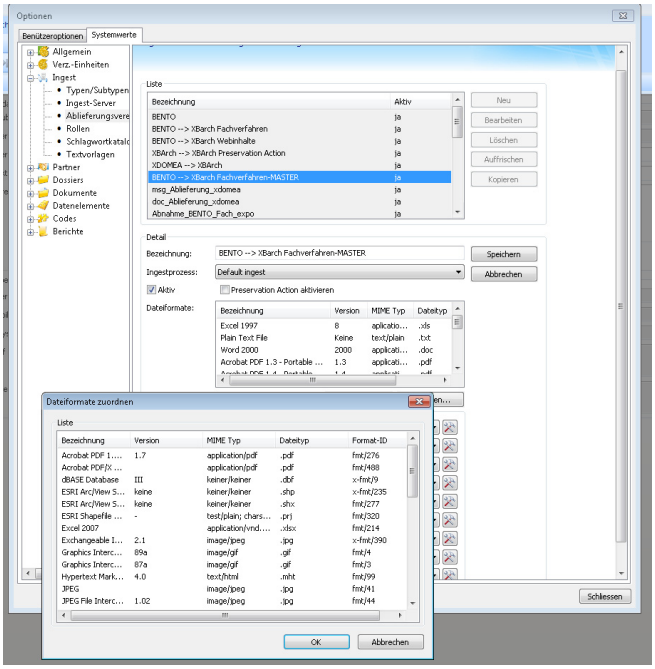

Figure 3, *Menu for ingest agreement*

If the SIP passes the first examination, more detailed investigations follow:

1. Data formats in the SIP are compared to the formats in the ingest agreement
2. The integrity of each file in the SIP is checked by comparing new MD5 hash values with submitted ones
3. The completeness of the SIP is checked by counting each file in the package
4. Each file gets a virus check

The common tool DROID is used to recognize the data formats in the SIP.

After that procedure, the SIP is unpacked and ready for the next important step.

Building the AIP

Each AIP that goes into the archival storage has the same basic structure

There are two directories. One for the metadata and one for the data objects. Every type of information required by the OAIS for a complete preservation description is located in the xbarch_metadata xml-file. Each xbarch_metadata xml-file consists of three sections. The first section contains administrative data about the creation of the SIP, the date of the data transmission and ingest.

The second section is a technical description based on the PREMIS schema. That's where the representation information is located. The result of the former format recognition by DROID is also documented in this section. The recognized format, represented by its Pronom-ID (PUID), is the link to a brief

formalized description of the hard- and software environment needed to visualize the archived object. The descriptions can be very short if the object is just a file in PDF/A format or TIFF. But if the object is a very special file format, the descriptions can be quite complex. An esri-shape file serves as a good example, delivered by a database that holds information about mining companies and their claims in Saxony. These complex files, which contain geographical data, vector graphics and a database, require a special viewer to visualize. This shows that we need additional information to describe the required hard- and software settings.

The technical description holds also metadata about integrity. Each file has its own MD5 hashtag, so users and archivists can check the integrity if they want to be sure.

If a migration has occurred during ingest, documentation is made as to which file was the source file, which converter did the job, and which file was the result of the migration event. In OAIS, this type of metadata is called provenance information.

Other important information in the second section is the field of significant properties. Due to the fact that digital objects behave differently in changing environments, and migrations can alter the whole look and feel of the objects, significant properties have to be defined at the very start of the archiving process. They serve as a benchmark to measure future preservation treatments. To define the significant properties, the following points have to be considered:

- The "look and feel" of an object before archiving starts
- The designated community
- How will the object possibly be used in the future by its designated community

Even though significant properties aren't part of the original OAIS terminology, the term is closely related to the designated community. The designated community in OAIS is important, because the community knowledge base shapes the content of an AIP, which must contain all information necessary to turn a data object into meaningful information. You can add less information to your AIP in case the knowledge base of the designated community is wide. State archives must assume that any citizen can claim a right to work with archival holdings. Therefore we have to add as much information to the AIP as we can.

The third section of xbarch_metadata.xml describes the informational content of the objects. In OAIS terminology, this is the reference and provenance information which completes our preservation description information. The structure of this section can vary with the type of the objects, but in general, this part of xbarch_metadata.xml answers the following questions:

- Which agency created the object?
- When was the object created?
- What is the informational content of the object?
- Which e-government system created this object?
- What type of object is it (e.g. a website, a database, a video, a sound recording or an electronic record)?

We don't save any metadata about rights in our AIP.

Different Objects – Different Ingest

As mentioned before, our system allows for three types of ingests depending on the incoming objects. The second type of ingest is made especially for electronic records. Records management systems were the first systems to catch the attention of archivists. And archivists took part in writing the first standards for records management at a very early stage. What MOREQ is in Great Britain and Europe, DOMEA is in Germany. DOMEA defines global rules for electronic recordkeeping, including an interface for the appraisal and ingest of records.

In our project, the plan is to implement the DOMEA interface both in the records management system of Saxon agencies and in the electronic archives. The "archives" part of the interface has been working since 2013, the records management system will follow in 2014.

With both sides working, we will be able to appraise and ingest electronic records in automatic processes. Messages between the two systems will be send in XDOMEA-XML Packages. The basic steps of the process are as follows:

1. The records management system creates an appraisal list for all records that are set to leave the agency.
2. The appraisal list is sent to the archives.
3. The archivist opens the appraisal list with our own program called XDBewerter and marks each record with an appraisal flag (thumb up, thumb down).
4. The list goes back to the records management system. The system places all records with archival value into a XDOMEA-SIP. Records on the appraisal list without archival value are destroyed.
5. The archivist must prepare the electronic archives and define an ingest agreement (as mentioned). In this case the ingest process expects an XDOMEA-SIP instead of a BENTO-SIP.
6. The SIP will be uploaded by the electronic archives and runs through the same ingest steps as packages of other types.
7. AIPs of the same basic structure, consisting of an xbarch_metadata.xml and data objects, are the result.
8. Upon completion of the ingest process, the electronic archives sends a mission accomplished message back to the agency.

The third type of ingest is part of the functional entity preservation planning. If the archivists decide that certain AIPs need to be migrated into another format, selected AIPs can be re-ingested.

Preservation Planning and Data Management

We upload the whole xbarch metadata into our database. The technical information is the basis for the functional entity of preservation planning. Again, the PRONOM-ID (PUID) is the main key to control the variety of data formats in the repository and to plan further preservation treatments. Information about the required technical environments for accessing our archived objects is also very important. If there is a high risk that some kind of environment will not work properly on a new generation of technology, for example a new operating system, the archives can plan a reasonable preservation strategy. Reasonable means, we first look if some minor changes on the recent environment can solve the problem. Perhaps all we need is a new viewer application, which is able to visualize the data object and which runs properly

on our respective operating system. If that solution fails, a major change such as emulation or virtualization could help. Migration should always be the last strategy, because it consumes time and storage, changes each data object completely, and has always consequences for the look and feel of the objects.

With the upload of the informational content in xbarch metadata, archivists are able to search and retrieve any AIP in the electronic archival system. The system has a simple search over all indices, but you can launch complex queries as well.

The content information in the electronic archives is not a complete archival description like an inventory. Thus, we will upload the informational content metadata into the main archival information system of the Saxon State Archives for further refinement (description) and to have one single "point of truth" for searching the archives. Every object in the holdings of the Saxon State Archives, be it paperbound or electronic, must be described in one database.

Archival Storage

Every AIP, having finished the ingest process, is transferred to the repository. This step represents the functional entity of archival storage. Storage is placed in two different locations within the borders of Saxony. Before storage, the AIP content and structure are transferred into one tar file. Each AIP is maintained 4-fold: three copies on hard drives and one copy on tape. The software running the repository is also very advanced. In principle, it could run an electronic archive by its own. Therefore, the repository software lets an AIP run through a second ingest, this time to proof the quality of an AIP.

Access

Access is the main reason to run an electronic archive. But the OAIS says nothing about how easy the access has to be. For most projects in Germany on digital preservation, ingest was the first challenge they had to face. Because of that, the concept concentrated on the transformation from SIPs to AIPs, not so much on the transformation from AIPs to DIPs. Ingest has its own DIN standard in Germany (DIN 31645), but there is no standard for access. For the time being, if someone retrieves an object from the Electronic Archives, the DIP he gets bears no difference to an AIP. That is not a violation of the OAIS standard, but most of the time it is inconvenient. If the user orders an AIP, consisting of standard formats like PDF/A or TIFF, he can access the whole content immediately in one browser-based viewer. AIPs with special data formats, like high definition videos, require certain technical environments, for example fast hardware and high definition screens. In that case, the correct technical environment has to be set up manually, according to the technical environment description in the AIP. So: Is it possible to access the whole content stored in the Electronic State Archives? Yes it is. Is the access fast and is it fun? Well, not yet. That leaves much room for development in the future.

Organizational Concepts

The second cornerstone is a reasonable organizational structure to get the work done. With the very beginning of the project it was clear that the Saxon State Archives can't do it without help. So the hosting of all servers and storage systems is left to the datacenter Staatsbetrieb Sächsische Informatik Dienste. The support of the main software is part of the contract with the software vendors.

New tasks for the archives are:

- Pre-ingest measures
- Ingest monitoring
- Preservation planning
- First-level support

The Saxon State Archives installed a new organizational unit, called the "Leitstelle", with four employees to deal with this increase of tasks.

Pre-ingest Communication

The project worked out a workflow to pass the needed information to appraise and ingest data from e-governmental systems.

1. Archivist notices a system in his agency containing data with potentially archival value
2. Archivist informs his colleagues at the "Leitstelle" by e-mail or by means of a special database (Vorfeldtracker)
3. The archivist documents the significant properties he wants to preserve
4. The Leitstelle defines a strategy for ingest and preservation planning
5. After a few ingest test runs, archivist and Leitstelle discuss the results
6. If both sides agree that the test was successful, the real ingest can start.

It's very important to keep the communication channels open between all parties involved.

Conclusion

In three and a half years (August 2009 to February 2013), the new Electronic State Archives (el_sta) went from a theoretical concept built on the foundations of the OAIS (ISO 14721) to a running infrastructure, able to archive e-government data from electronic records to web-content. But an electronic archives is not rigid, instead it is constantly subject to change. It has to adapt to changes in technology, to user preferences in respect to access and working with data, and to the legal frameworks of e-government.

References

[1] ISO 14721:2012, Space data and information transfer systems – Open archival information system (OAIS) – Reference model

[2] DIN 31644, Information and documentation – Criteria for trustworthy digital archives (Beuth Verlag, Berlin, 2012)

[3] DIN 31645, Information and documentation – Guide to the transfer of information objects into digital long-term archives (Beuth Verlag, Berlin, 2011)

[4] DIN 31646, , Information and documentation – Requirements for the long-term management of persisten identifiers (Beuth Verlag, Berlin, 2013)

Author Biography

Karsten Huth graduated at the Humboldt University in the fields of library science and linguistics (2004). He worked for the German Federal Archives as an expert for digital preservation (2005-2009) and proceeded to the Saxon State Archives, where he leads a project tasked with the goal of building an electronic archive system.