Ways to deal with complexity[1]

Christian Keitel

Landesarchiv Baden-Württemberg
Staatsarchiv Ludwigsburg, Arsenalplatz 3
71638 Ludwigsburg, Germany
christian.keitel@la-bw.de

Abstract

Several ways to deal with complexity are discussed. An archive can handle the matter by keeping the number of the single elements in the core areas of digital preservation down. The numbers of action types during the ingest process, of metadata and journals could be reduced. A preservation model for analogue and digital records is outlined. By keeping complexity down, it's easier to see what digital and analogue archiving have in common. Instead of seeing two totally different worlds (here is the old one, there is the new one), one can shift to a less revolutionary view. This makes it possible to fall back on the considerable implicit knowledge of the existing memory institutions. From the perspective of the whole archive, there are strong arguments for reducing complexity and keeping digital and analogue things together whenever possible.
Complexity can also be handled by cooperation. The Landesarchiv Baden-Württemberg appreciates the opportunity to use the software tools DROID and JHOVE. The BOA project is a further example for a venture in website archiving that is maintained by libraries and archives collaboratively.
Cooperation and the reduction of complexity are the two most promising ways to enable small and medium sized archives to start with digital preservation. Automation seems to be a good thing whenever it can be achieved, but until this stage is reached, the single steps and the standards which must be followed often are extremely complex.

Complexity matters

Over the past years considerable progress has been achieved in the area of digital preservation. PREMIS explains which preservation metadata should be kept; METS describes how to build an information package; PAIMAS lists nearly 90 steps for the ingest process and DRAMBORA enumerates the possible risks of digital archiving on more than 200 pages. These standards or guidelines have resolved many of the open questions. On the basis of these results and foundations, it should be easy to build a digital archive. Therefore it is striking that these achievments have not been followed by a significant increase in the number of digital archives. Although many memory institutions have assumed the task of securing and preserving digital objects, only some of them are actually doing this. How can this discrepancy between the progress of digital archiving and the widespread failing of implementations be explained? Has there at least been a public discussion of this problem?

Three observations may contribute to the search for an answer:

1. Beyond doubt, the named standards and guidelines are all extremely helpful. Their detailed information addresses both general and special problems. But it is a hard job to extract from these texts some general hints how to start with digital archiving. This task is even harder for a beginner in digital archiving.
2. The communities of the traditional archivists on the one hand and the digital archivists on the other hand are deeply divided. Each community is oblivious of the other. Hence, the implicit knowledge of a still existing memory institution is only rarely taken into account when setting up standards for digital archiving.
3. Standards are usually devised by members of big institutions like national archives or national libraries. Once again it must be stated that these are very valuable contributions. But are they equally applicable to smaller archives or libraries?

For many memory organizations, complexity is one of the most serious impediments to start with digital preservation. The extensiveness of the standards and the large number of articles published raise the suspicion among librarians and archivists that digital preservation is something nobody can really cope with, nobody or only the biggest memory institutions. It seems to scare all people who are supposed to establish digital archives but so far haven't started. But complexity is more than just a psychological problem. In the long term, complexity makes preservation more expensive and less feasible. So it is worthwhile to think about how we can deal with it.
One possible answer to this question is cooperation. Cooperation takes centre stage in many articles, projects and conferences. Although the necessity for cooperation can't be overestimated, it is not the only way to deal with complexity. Archives can also try to reduce it. For example, many specialists in digital preservation keep the number of their archival formats down. Thus, they reduce the complexity of digital preservation. But beyond this example there is remarkably little discussion about this option to deal with complexity. A third possibility to reduce complexity would be automation.

1 The paper was originally given at the iPRES conference, British Library, 9/30/2008.

Some of the recently published recommendations can be seen as a preparatory work for further automation. The preservation manager simply presses a button and all the complex work will be done by the machine. But defining such maschines seems rather complicated, as you can see, for example, at MoReq2. As a result, the recommendations are growing more and more complex while the implementations (the machines) are still out of sight.
Summing up, complexity seems to be a serious obstacle on the path to digital archiving. This paper describes some of the ways in which the Landesarchiv Baden-Württemberg tries to deal with it. The results presented below were devised in the course of the project "Digital Archive in the Landesarchiv Baden-Württemberg", running from the end of 2005 until 2009.

Standards on Ingest

The Open Archival Information System, better known as OAIS, describes six functional entities: Ingest, Data Management, Archival Storage, Preservation Planning and Access. Altogether, the standard describes about 30 functions. In the summary chart these are connected with each other by almost 70 (68) arrows. What does this mean for someone trying to set up a digital archive? Even if each arrow corresponds to only one task, there still is a lot of work to be done.
For the ingest area OAIS specifies the following functions:

- receiving SIPs
- performing quality assurance on SIPs
- generating an Archival Information Package (AIP)
- extracting Descriptive Information from the AIPs for inclusion in the archive database and
- coordinating updates to Archival Storage and Data Management.

The functions are characterised in a highly abstract way and they are not ordered chronologically.

Two years after the publication of OAIS, the Management Council of the Consultative Committee for Space Data Systems (CCSDS) released a second recommendation: The Producer-Archive Interface Methodology Abstract Standard. PAIMAS gives a more detailed view of the relationships and interactions between a producer and an archive. Although the specification covers only the first stage of ingest, it still needs 86 steps to describe the transfer of a record from the producer to the archive. This is divided into four phases:

- Preliminary Phase (46 steps)
- Formal Definition Phase (36 steps)
- Transfer Phase (2 steps) and
- Validation Phase (2 steps).

Speaking of "phases" implies a chronological order of the single steps. In fact, the recommendation starts with the identification of the contact persons and the exchange of general information (P-1 and 2). PAIMAS here is much more concrete than OAIS, but can the recommendation be understood as a true construction plan for a digital archive? There are at least two arguments against this assumption: Firstly, it seems nearly impossible to go through 86 steps just to run the first half of the ingest process, i.e. to transfer an object to the digital archive. Secondly, the concept lacks flexibility. The strict chronological order of the single steps forces the readers to go gradually forward. As each step is based on another, their order can't be changed. The catalogue of PAIMAS therefore needs further transformation to become a construction plan for the Ingest to a digital archive.

An interesting proposal was made last year by the members of the Australasian Digital Recordkeeping Initiative (ADRI). They designed a Submission Information Package. Deliberately, a number of questions are not addressed. Nothing is said about the high level transfer process or the low level protocols or the physical transfer mechanisms. In other words, ADRI has done two things: On the one hand, they concentrated on a decisive part of the ingest process, and on the other hand, the result of their work (the SIP) allows each institution a lot of flexibility.

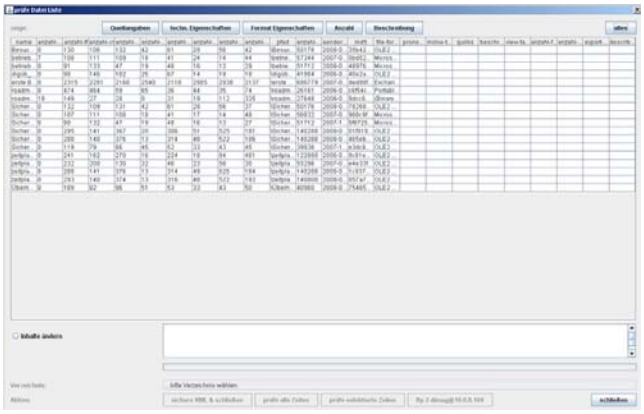
Ingest

Following OAIS, many institutions are forced to preserve their digital information in a way which is similar to the work of the traditional archives. Therefore, if a traditional, paper-based archive goes digital and plans to preserve digital information, many of the functions mentioned in OAIS are already well known. A traditional archive can thus refer to the implicit knowledge of its staff. So, it is not important to recall all the steps of PAIMAS. Anyway, if one takes into account the implicit knowledge of a memory organisation, prescribing a fixed ingest process seems to be rather the wrong way. Every archivist could name cases, in which the normal sequence of steps can't be maintained. On the other hand, it is obvious that the traditional ingest process is not sufficent for the transfer of digital objects. Hence, a fundamental analysis of the whole process was done during the project, aiming at maintaining both the flexibility for the archivists and the manageability of the ingest process. As a main result, a distinction between action types and process steps was introduced. An action type can be seen as a tool: You can use it whenever it is necessary in a single process step. One action type can be used in different process steps. Of course, there is a typical way to proceed in the ingest process, so a list of the normal sequence of the single process steps was drawn up. But it is important to point out that nobody is forced to follow them in the order listed.
How many action types should be distinguished? Within the context of a traditional archive their number can be reduced to four: Appraisal, inventory taking, transfer and validation. These are the action types which are essential for the Ingest of digital information.

Appraisal stands at the very beginning of the ingest process. It is the only action type that occurs at only one stage of the whole process. Appraisal can be divided into three parts: First one has to decide if an object should be taken into the archive and be permanently preserved. If the object is part of a large and not clearly delimited system, archivists have to define the boundaries of the object (e.g by specifying the tables of a big database system). They must also define which properties of the object are significant. After the appraisal it should be clear which object in which form and with which significant properties is to be preserved over time.
The other three action types are closely interrelated. They can almost be understood as a template. In its centre stands the transfer. The entire ingest process can be seen as a succession of transfers: to a new system, to a new data carrier and at least to the archive. But a transfer itself isn't enough. Each transfer means incertitudes about its results. For this reason it must be ascertained that the result complies with the expectations. This checking after the transfer is generally called validation. Validation can be seen as a comparison of two things: One object defines the desirable outcomes and the second should show exactly these values. Instead of an object one can speak about the specific properties of the object. These properties can be described in inventories during the Ingest. It should be noted that at least a part of these properties are identical with the significant properties mentioned above. Validation therefore is the third, inventory making the fourth action type.
A case in point: Let us take a big database system as an example. Eight tables are to be taken into the archive, with each table becoming one CSV file. What are the single steps during the ingest process?

1. Appraisal: The archivist has decided to preserve the information and selected the eight tables. He or she must define at least the significant properties: The sequence of characters within each field should be maintained, the links between the tables and so on. Some of these properties are countable, e.g. the number of fields.
2. Inventory making (1): Some properties are gathered in the database system: Number of tables, fields and datasets. They are written in list 1.
3. Transfer from the database system to the CSV files (migration).
4. Inventory making (2): The same properties as in step 2 are gathered from the CSV files.
5. Validation by a comparison of the two inventory lists.

If the validation fails, the process has to be repeated. In case of success the next transfer (to the archive) can be prepared. Note that even if there is still a valid inventory list, some properties can be collected only now: Think, for example, of a hash value of each file. But in general, the process can be repeated for the transfer to the archive, the transfer into the repository and even during a future migration of the archived files: Inventory taking, transfer (migration), second inventory taking and validation.
The order of steps which has been described is the most common one, but it can be modified if necessary. Sometimes several transfer processes within the producer area are necessary. Sometimes the appraised objects can be transferred immediately to the archive. But whatever sequence of steps will be chosen: The four action types are enough to develop the ingest process step by step.
PAIMAS, of course, includes some steps which can't be processed by these action types, e.g. steps associated with the legal circumstances. But these aspects are not specific to digital objects. Here we can count on the implicit and explicit knowledge of the archivists.
During each ingest process, several lists with properties of the appraised objects are assembled. We decided to collect these data in one software tool. IngestList enumerates the single files with their core data like "file name", "date saved", "MD5 value" etc. A DROID-Integration allows the identification of the Format-ID of PRONOM and other criteria, too. More properties are taken from JHOVE, which is also integrated. Also, the tool gathers the most common values of field delimiters and dataset delimiters of the CSV format, e.g. control, line-feed, pipe, semicolon etc.
All these lists can be easily compared. Hence, validation is the second task of the tool: IngestList compares the lists and establishes both consistency and discrepancy. For each ingest process, all information gathered is inserted in a single XML file. This file is accompanied by a MD5 file which makes it difficult to falsify the content of the XML file.



IngestList

With every ingest process, a gap has to be bridged between producer and archive. This gap starts with the appraisal, often contains a migration and ends in the OAIS section Archival Storage of the archive. After the process, the records should be as trustworthy as before the ingest. Therefore, all activities during this phase should be documented. The more we know about the actions and circumstances of this phase, the easier it is to claim that the records are trustworthy. With its lists, IngestLit helps a lot: They contain so much information that it would be difficult to change any value without being noticed. But the tool documents more than this. IngestList also contains a special journal section, where all single steps or actions can be entered. Some

information is taken automatically, e.g. who has done the inventory making or a validation, when it was done and what the results were. Some information can be added by the archivist, e.g. why a single step was taken or how an export was done up to the insertion of a SQL statement. According to OAIS the ingest process ends with the transfer of the objects to the area of Archival Storage. At this moment, IngestList contains a full journal of the entire process. Due to the MD5 file and the amount of related information, which allows many cross checks, the information contained in the journal as well as the single lists give good evidence about trustworthiness. At the same time, IngestList doesn't require a fixed sequence of single steps. Therefore, the archivists are as flexible as they are with the Ingest of paper records.

Archival objects

Traditional paper based archives only work with one kind of objects: A paper record comprises the logical information and the physical carrier in an inseparable way. Each object has its defined limits.

But things are different in the digital world. PREMIS has shown us the fundamental split between logical and physical objects. According to this standard, the latter fall into three subtypes. Digital preservation itself seems to be much more complicated than the preservation of analogue materials, and the two tasks seem to be completely separated. But on the other hand, with microfilming and digitising of analogue objects we have already crossed the boundaries of the analogue world. Therefore, some important questions that came up were whether all kinds of records could be unified in one system of description and whether this could be done in a fairly simple way.

PREMIS distinguishes between representation, file and bitstream. A representation embodies the logical information (intellectual entity) and can contain files and bitstreams, whereas a file can contain just one or more bitstreams. Hence, a bitstream must depend on a file, but a file can depend on a representation or immediately on the intellectual entity. So, one entity of the analogue world is opposed to four objects in the digital world.

First, let us look at the world of the digital objects. The representation allows us to name exactly that bundle of files which represents a record. For this reason it is obvious that in many cases we need the concept of representation. But the question was: Is it acceptable that some files depend immediately on the logical object and others are part of a representation? Making use of representations means preserving different versions of a digital object over time. Making no use of representations in the case of a migration means either overwriting the old file or renaming the new file. Is it possible to preserve millions of files over centuries, some of which with their predecessors preserved, others without them, and still others bound together within a representation? To adopt this model would increase the number of different preservation paths and therefore also the complexity of future decisions on preservation. So, in this case we argued against flexibility because we didn't want to allow totally different preservation paths. All our digital records therefore have at least one representation. The digital representations consist of files, but a file can't depend immediately on an information object.

Our second question was: Is it possible or even recommendable to introduce the representation model for the analogue born objects as well? Obviously we live in a time of copies. If you want to preserve some of the copies of analogue materials for a long time and make them searchable you have to think about the representation model. For these reasons, the Landesarchiv Baden-Württemberg has decided that all records (digital and analogue) should have at least one representation. Both representation and intellectual entities are listed in our finding aids (the OAIS area Data Management), whereas the digital files aren't shown there. So, analogue and digital materials are described in the finding aids together and in the same way.

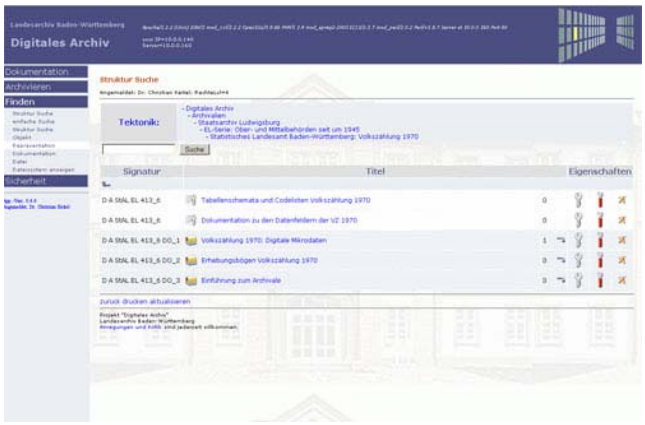
The representation model presented above allows us to start our preservation activities for all kinds of objects in the area of Data Management. Some of the analogue born objects have to be preserved (e.g. a parchment charter), others can't be preserved (e.g. a drawing in glassine). Many of them are listed alongside another representation. Seeing a logical object with more than one representation means seeing different opportunities for preservation. Thus, preservation planning can almost be seen as information management. It has to be stated, though, that the material properties of a medieval charter in this context are a part of the information as well. However, there is a common entry point for all archival objects in our system, and the number of preservation paths and the complexity of the preservation have been remarkably reduced.

Archiving system

Looking for a repository system can cause severe headaches. First you have to define your requirements. Which objects should be archived? There is a great diversity of digital object types which archives may want to preserve. Most of these objects are embedded in hierarchical structures, which are not standardised but quite flexible. See for example the classification schemes and their distinction between series, files, subfiles, records and documents as described by MoReq2. These structures should be preserved together with the records themselves, but it's not easy to define the exact borders of each object. As a result, we have to maintain different objects complete with the logical links between them.
Another requirement was to keep the Archival Information Package within the file system and to use a database system with redundant metadata information only for management tasks. This means that the AIPs must be exportable from the file system even if the repository software fails or can no longer be operated. On the other hand, this possibility should be open for the administrator only. For the archivists there should be only one entry point (the repository itself) to the AIPs, coupled with a user management system. In 2006, none of the repositories inspected was able to meet these requirements. So we decided to build a new one which suits the requirements of an archive. Needless to say,

"archive" here means the traditional memory institution.
If you decide to build a new repository on your own, the headache is even growing. Is it possible to construct a repository for all kinds of digital objects or should there be one repository for each type? Presumably, many archivists and librarians would opt for the "one fits all" solution. But in practice, differences can be noted: At the moment, some archives concentrate entirely on only one object type. They're working on a fully automated import function and a suitable repository. Other objects are expected to come into the same repository later, but the practical plannings for this are postponed.
This is a common strategy of traditional archives: They are looking for solutions for current digital records. Of course, it is important to save this information. Concentrating on one type of objects is also a way to reduce complexity. But at the same time other potentially important records like e.g. databases are neglected. Therefore, we've decided to build a repository which from the very first day can import all types of digital records. A metadata model, which covers about three dozens of core metadata, stands in the heart of this repository. Dataset-ID and file-ID, signature, title, description, provenance, time, state, creator and others are collected in a structured way. Many of these can be captured automatically. For each record type, other structured data can be implemented. In the case of databases, there are fields for the number of datasets or columns. Furthermore, non-structured metadata or documentation can also be used. Documentation is always welcome, but except for the above mentioned structured metadata we don't make an effort to fill each logical information unit in its own data field.
The combination of core metadata, expanded metadata and documentation makes it possible to define only one way of transferring records into the repository. Each object could be sent into the repository manually. But with IngestList it's possible to transfer them automatically.
Another important feature is the protocol function. It is clear that a repository comprises a lot of duties. Many of these should be listed in a journal so that a future user can consult them in order to verify the trustworthiness. But if each task produces a journal of its own, this would result in a mass of information in a lot of different places. Therefore, we've decided to bring all the valuable protocol information together in two kinds of journals: One for each AIP and one for the archive as a whole. These journals aren't log files, and they are not written in some proprietary file format as it is important to keep them readable for the near and the remote future. For this reason, both protocol types are written as XML files. Each one has its hash value so that it's difficult to change them without being noticed.
The reduction to a small number of metadata, no more than two protocol types and only one way into the repository helped us to keep the complexity down. As a consequence it was possible for the project team to develop and to implement the digital repository DIMAG with just three persons in 2006.

DIMAG stands for Digitales Magazin (digital storeroom). It is able to hold all kinds of digital objects. In 2008 the digital repository comprises more than 18.000 digital records, including databases, pictures and textual records. Due to the reduced number of metadata fields and protocol types it is not complex to handle DIMAG.



DIMAG

As previously mentioned, DIMAG is able to handle all types of digital objects; the Landesarchiv Baden-Württemberg keeps nearly all archival objects in it. But there is an exception to every rule. Although it would also be possible to keep websites in DIMAG, these are preserved in BOA (Baden-Württembergisches Online-Archiv). This system is run by the Bibliotheksservice-Zentrum (Library Service Centre) Baden-Württemberg (BSZ), a support institution for libraries and archives. The two state libraries and the Landesarchiv cooperate in the archiving of websites in this system. In this case, complexity was reduced by sharing the risks with other memory institutions on the basis of a common object type; in other words it was reduced by collaboration.

Acknowledgements

Rolf Lang has programmed and implemented IngestList and DIMAG. Kai Naumann has taken the majority of objects into the digital archive. The author thanks these two colleagues for their invaluable contributions. Special thanks also to Heidrun Wiesenmüller for further discussion and suggestions.

Notes

Further information can be found at http://www.landesarchiv-bw.de. Use the full text search (entering "DIMAG") or see under "Fachinformationen" >>>> "Elektronische Unterlagen".