Digital Archaeological Resources at the University of Bergen: An Efficient Tool in Research and Heritage Management?

Asbjørn Engevik jr[2], Jon Holmen[1] , Sonja Innselset[2] and Jarle Stabell[1]

[1] The Norwegian Museum Project, Oslo
jon.holmen@muspro.uio.no
jarle.stabel@muspro.uio.no
[2] The Norwegian Museum Project, Bergen
asbjorn.engevik@ark.uib.no
sonja.innselset@bmu.uib.no

Abstract. This paper describes the process behind the creation of a digital version of an archaeological archive at Bergen Museum, the University of Bergen. The original archive was scanned, a data model for the information structure in the documents was defined and information from the archive was entered into a database based on this model. Examples of usage of the digital version in the cultural heritage management are also given.

Keywords: Digitisation, databases, information retrieval, data modelling.

1 Introduction

In this paper we introduce the digital archaeological resources at Bergen Museum, the University of Bergen. First we present the original archives in the museum, then we go on to outline briefly how we went about the digitisation process, and then we explain the database solution in detail. We will also demonstrate how these electronic resources are, and will be, efficient tools in research and cultural heritage management. We finish by outlining possible future refined extensions of these resources.

2 The Documentation Project at Bergen Museum

This digitisation project is a part of the Norwegian Documentation Project, now called the Norwegian Museum Project. In the course of our project several resources have been made accessible in a digital version, among them the three major archaeological archives at Bergen Museum:

- the published ancient monuments survey
- the published find catalogues and
- the topographical archive

Before we proceed any further we need to explain the term "topographical archive", as the usage in Norway differs from the common usage elsewhere. Our topographical archive, or perhaps we should rather call it a collection of documents, contains all written documentation relating to archaeological finds and ancient monuments in a specific museum region. It is called topographical because the information is systematised according to administrative geographical divisions: county, municipality and farm.

3 The Archaeological Resources

The first archive to be digitised was the ancient monuments survey. This comprehensive survey was published in the period 1950 to 1975, one booklet for each parish in the 77 municipalities in Western Norway, approximately 5000 pages.

The next archive to be digitised was the published find catalogues, about 4500 pages. These two archives were transferred to electronic text and tagged in SGML. They are now linked, freely accessible on the Internet, and it is possible to make combined searches in them.

The topographical archive is the third and last one, and the most extensive. But unlike the two former resources it will not be accessible on the Internet without authorisation through a password because it contains several references to private issues. In a way, the two former archives are offsprings of this one. This archive contains all existing written documentation about the various artefacts in the museum collection as well as about the ancient monuments in Western Norway: letters, field survey and excavation reports, drawings, maps, etc. This archive was established in 1825, or rather the contents date back as far as that, to the founding of the museum. The topographical archive is in daily use, and new documents are added regularly. There are now altogether about 105,000 pages in 55,000 documents.

All these archives are in frequent use by the museum staff and graduate students, but searching for information is time-consuming and slow. Important information is easily missed and you often spend time looking for information, which is not there, so a more efficient way of extracting this information would be highly welcome. Another aspect is that many documents are very old and fragile, the paper is crisp, the ink is fading and the writing becomes illegible. With each subsequent handling, the documents deteriorate further. It is therefore important to be able to use these resources without actually handling them.

The answer to both these problems was to make an electronic version of the archives, to create a combination of databases and scanned documents. Thus both preservation and research/management issues are taken into consideration.

4 Overview of the Work

In brief, we have made an electronic copy of the entire archive. First all the documents were numbered and scanned as 1-bit pictures, in a multipage format. Then the most important information in the documents was registered in a database application. Different types of information are focused on. What kind of document do we have? Is it a letter, a report, a note, an application or a newspaper article? What is the subject? Who is the author or sender? To whom is the letter addressed? Are there any photos, drawings or maps included in the document? If the letter or report is about a find or an ancient monument, it is important to include the identity number. What kind of activity or event has brought forth the document? Was it a field survey, an excavation, was there a conflict between modern development schemes and prehistoric monuments, has any damage been done to an ancient monument or has a new discovery been done? All of these aspects, which in fact cover the entire activity range at an archaeological institution, are present in this electronic version of the archive.

Many documents are related to the same event. For instance, there might be several letters, notes, reports etc. from the same excavation or field survey. These individual electronic documents are now linked through a common case number.

5 The Scanning Process

All documents have been scanned and stored as raw image files. We chose 1-bit multipage TIFF, one file for each document, as storage format for the scanned documents. The main reason for choosing a 1-bit format was the size of the archive and the need to save disk space.

Looking back, one can say that the one-bit-solution was only 90% successful. We did a lot of testing before purchasing a suitable highspeed scanner, but later we found that about 10% of the documents needed greyscale to be readable at all.

When scanning old documents, which are not quite black and not quite white, one has to do a lot of adjustment in the scanner settings to obtain readable images, so we lost much time and quality in that process.

What we also learned from this is that imposing limitations on quality based on hardware limitations should be done with care. In this case the cost of storage has been dropping all the time, and made grey scale scanning possible.

6 The Data Model

The data model, which has been more or less unchanged during the whole period, is basically divided into two parts:

- One concerning the organisation and storage of the real (scanned) documents, *the physical archive data model.* This part was more or less straight forward.
- One concerning the extracted data from the documents, *the extracted data model.* That part was not at all that straight forward, but reflects what we wanted to get out of the documents.

6.1 The Physical Archive Model

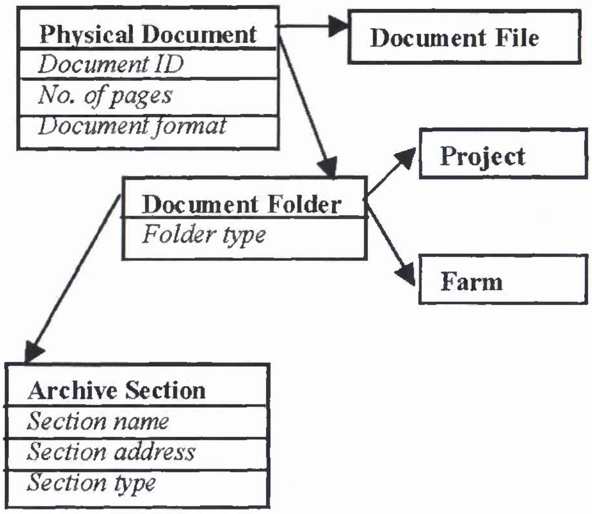

Fig. 1. Model for physical archive

The physical archive model is nearly a copy of the real archive organisation with the *archive section object* as top level. Archive sections keep track of which drawer contains which document folder.

The archive section is referenced by the *document folder objects*, one for each folder in the real archive. It is either of project or farm type. If it is set to project, the folder id links it with a specific project, normally an excavation project, if it is set to place, the folder id links to a farm.

The document folder is referenced by the *physical document objects* of all documents in that folder. The physical document has a property called the document number, which is also marked on each paper document, so that one can easily check whether the record in the database matches the scanned file. It also contains information about the document format, that is, TIFF, Word, etc.

For the time being this property is redundant since all documents are TIFF 1-bit Multipage, but new documents could come in other formats such as Word, raw text, SGML-tagged, etc. It is also possible to access the scanned file that is stored somewhere on the operative systems disks through the physical document object.

6.2 The Data Extraction Model

The model for the data extraction part is somewhat more complex.

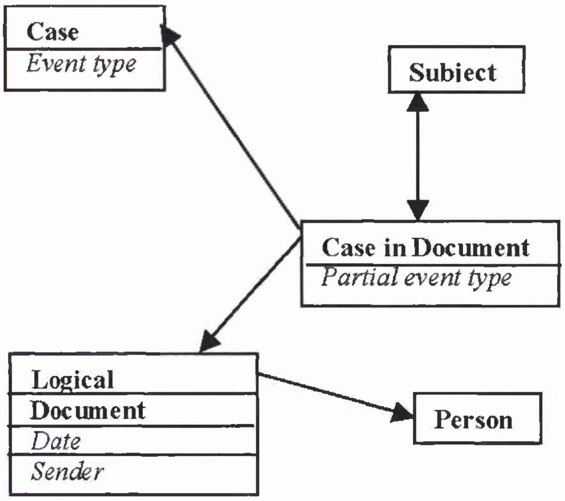

Fig. 2. Model for extracted data; logical document vs. cases and events

Three obvious things we wanted to extract were *date*, *sender* and *recipient*.

We discovered that some of the old documents looked rather like new e-mails that had been sent and re-sent a couple of times, with several messages and replies in the same document. This led us to introduce an object-type called *logical document* that captures each of these single mailings.

Each logical document references the physical document it is a part of, and from each of these the date, sender and recipient is extracted. We are thus also able to build a person gallery from the museum's activity.

The logical document has another property called document type to identify whether it is for instance a letter, a report, a note, or a newspaper cutting.

We also wanted to extract the issues or *cases* treated by the museum during the archive period. This is what you might call the archaeology of the museum's activity.

To do so the documents had to be read thoroughly. Again we found that within each logical document several different cases or issues could be mentioned, and that the same case was generally mentioned in several documents. We decided therefore to introduce another object type called *case_in_document.*

Each of these case_in_documents references a *case object* with the same id.

The algorithm of getting this thing right is a kind of reversed engineering:

- Identify individual cases inside each document.
- Check if any of the cases are parts of a case mentioned in another document and if so, give them the same case number.

The case object holds the whole case or issue. It has a property called *event_type*, which can be excavation, registration etc.

The cases_in_document also have this property, and capture what can be called the sub-events. So by searching the case_in_document-table for a specific case and sort the hits on date, you get the sequence or thread for that specific case.

The *illustration types* and numbers for each case_in document are also extracted together with a *subject.*

This subject is more a short description suitable for keyword-search on what the specific case_in_document is about.

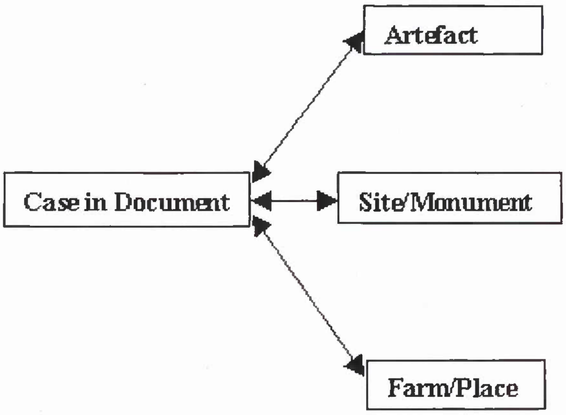

Fig. 3. Model for extracted data; case in document vs. archaeological objects

Even if the topographical archive is organised according to the farm that the documents covers, cases tend to stretch over more than one farm. So we also had to extract all the *farms* treated in each case in each document.

By doing so one might say that the electronic version, with respect to anything but where the real document is stored, has made the old farm-folders redundant. It is far more interesting to search for which farms are involved in a case than for which folder it is stored in.

We also wanted to extract the *artefact identity numbers* and *monuments* that were mentioned in each of these case_in_documents so that it would be possible to link the documents to the museum inventory databases and the sites and monuments registers.

7 Taking the Application into Use

As we see it, the electronic version of the archive is, or rather will be when it is taken into use, a very valuable resource in both research and heritage management. We will give just a few examples to demonstrate how easily and quickly relevant information can be extracted. Another important aspect, in addition to finding what you are looking for, is the amount of time saved by finding out, in a matter of seconds or minutes, that the document required was never included in the archive. Or perhaps that it was removed before digitisation started. This also means that one can do queries which were practically impossible to do in the non-electronic archive. We will give a few examples of some of the simple, but important searches we can execute.

7.1 An Efficient Search Tool - Examples

One very common question is whether there are any documents (or rather any activity) in the archive relating to a specific farm. This question arises every time a new road or any other development project is planned. We start by selecting the relevant farm. If desired, we can choose to sort the documents by date, and thus get a good picture of the temporal activity on that particular farm. To find out more about a specific document, we can choose to view the scanned image by clicking the selected document twice. From there the document can be printed for further scrutiny or use in the field. There might also be references to this particular farm in documents elsewhere in the archive. This is the kind of information that you will not find in a manual search operation in the archive, but here it follows directly from the extracted farm information in each case in each document.

In the field ”Subject” it is possible to execute freetext searches. For educational or research purposes this is particularly relevant. A search for *dagger* reveals all documents with “dagger” mentioned in the subject. If you want to find out more about the find of a particular dagger, you just select that document and search for related documents in this specific case/event.

You may want to find all newspaper cuttings in the archive, in which case you select the document type ”Newspaper cuttings” and get all those hits from the archive. These cuttings are always part of a wider case, and you can choose to view the rest of the related documents.

These three searches are all very simple searches; only one criterion was used at a time. A search can also be narrowed down by choosing several criteria. For instance, if we are interested in excavation reports by a particular archaeologist, then we can select the document type ”Report”, the event ”Excavation” and the name of the person.

7.2 Statistical Analysis Possibilities - Examples

The efficient simple search tool is important in everyday use, but there are further important search facilities of particular interest. The possibility of grouping and comparing information across the entire archive, and the possibility of linking information to the other digital resources, opens up nearly unlimited opportunities for finding facts about archaeological finds, excavation and surveying intensity, amount and rate of damage to prehistoric monuments and sites, personal profiles, etc.

- If you want to find out about the archaeological activity in a region, for instance in the period 1955 to 1975, a search for field survey reports from the three different counties in that region will show you whether there is any difference in surveying intensity among them. These figures can then be compared with the corresponding excavation figures from the same period, and with the total number of new finds. The result can be useful for statistical purposes, or it can be used in a broader research on heritage management problems. An interesting research question could be what causes the difference in activity.

- It is a fact that a considerable number of ancient monuments, particularly grave mounds, have been illegally removed in this century. By executing a search for police reports for a specific period, it is possible to get an overview of how many of these cases were tried in court, and the outcome of the trial.

- From a management angle it could be interesting to find out the number of excavations in the last 50 years. This is easily found by searching for these excavation reports. This search can also be refined, for instance to find out more about who was responsible for the dig, how long it did last and how long it took to finish the report!

8 Future Extensions

One important future extension is to integrate the topographical archive and other archives with the registers for artefacts and sites/monuments.

The keys we have extracted from the archive can be used to connect with the records in the other bases. Thus when you search for information about a specific artefact, you will get hits on all related documents, that is, photos from the photographic archive, excavation reports from the topographical archive, etc., regardless of which physical archive it belongs to.

Seen from a focus on artefacts and sites/monuments research, it is not always important to know to which physical archive the document belongs. Generally speaking, one might say that when paperbased archives are converted into electronic databases, the old paperbased organisation of these archives becomes less important.

9 Conclusions

As mentioned earlier this application has not yet been taken into use by the staff at the museum. It has, however, been described and explained to them on several occasions, and the expectations are very high. The most obvious challenge now is the human factor. Will we succeed in making them use the electronic version instead of the original paper? This depends on the availability of the application; each user must be able to open it in his/her office and the interface must be simple. Success will also depend on the accuracy of the information in the database, whether it will be possible to feel confident about the search results or not. The complexity of the database necessitates a certain amount of background information, and the users must be given a short introduction on how to use it, in addition to a short user's manual.