Hybrid Architectural Archives: Creating, Managing and Using Digital Archives
(NAi Rotterdam, June 10-12, 2009)
Session on “Acquiring and processing digital archives in repositories”

Connecting the digital with the physical LAM: building a digital repository for the NAi

Henk Vanstappen
Project manager CIS
Nederlands Architectuurinstituut, Rotterdam
h.vanstappen@nai.nl

Introduction

The Netherlands Architecture Institute holds a vast architecture collection, including a library of over 40.000 books and eighteen kilometers of shelves containing drawings, sketches, photographs and other materials. Along the archive and library, the collection of architectural models and the nearby Sonneveld House with its contemporary furniture collection are typical museum-like collections. The NAI is entrusted with the safekeeping and management of these archives and collections and with making them accessible to the public. Researchers, students, and anyone else interested may consult publications and archives in the reading room.

These collections are all interconnected: a book may have been part of a architect’s archive that has been acquired, and the same goes for architectural models. Drawings that are part of a archival file may be exhibited in house or elsewhere, and so gain the status of a museum *object.*

Every item in the NAi collection is intellectually connected with an architectural project, an architect and/or an architect firm. A project can have been part of an event, such as a competition or an exhibition. An exhibition may have been about an architect and may have been documented in a catalog that's being held in the library, and so on. Information on people, organizations, projects and events is thus shared among library, archive and museum items as so called *authority files.*

To manage this LAM[1], the NAi is developing a collection information system (aka Collection Information sysyem, or *CIS*), based on the Minisis software[2]. The system allows us to register every item with its adequate standards[3], while maintaining the relationships between them.
The CIS also supports and documents actions or procedures that are connected to the management of a collection: acquisition, loan in, loan out, exhibition, movement, conservation, deaccession and so on. These are procedures typical for a museum context, and covered by the SPECTRUM standard[4]. The system also provides an OPAC where end users can search all the collections together or separately.[5] A more

advanced web interface will be bult in the Archivista project (2009-2010), which will have more advanced search and browsing options, web 2.0 functionality, user friendly navigation etc.

While the CIS was initially set up to manage analog, 'real world' objects, the number of digital objects in the NAI has grown substantially over the last years. First of all, digital reproductions of archival material being ordered by publishers or researchers has resulted in some thousands of files, that are stored on a file server. As demand for this material is still rising, this number is expected to grow exponentially. Next to this 'digitization on demand', the NAI has started systematically digitizing archival collections. The first project of this kind is the digitization of the J. Duiker archive, a project sponsored by Senter Novem (2009-2010). A third source of digital objects is the acquisition of new architectural archives that hold born digital objects. Since the NAI acquires archives mainly from architects who retire, the share of born digital material was minimal until recently. The Carel Weeber archives will be the first acquisition with a substantial share of born digital material.
With the expected rise in digitized/born digital objects, the NAI faces a new challenge: how will all these bytes be managed and preserved, and how will all this be connected with CIS and Archivista?

Survey
A first step was to acquire knowledge about the digital material that was available now or in the near future. Associates from four architectural firms were interviewed on their policies regarding use, storage and archiving of born digital material. Major conclusions of this survey where:

- All firms work in a Windows and/or Mac environment
- Software packages in use are diverse, but mostly mainstream: AutoCAD and VectorWorks for drawing, Adobe products for presentations[6], and Microsoft Office is popular as an office suite - but so was WordPerfect.
- Transition to newer platforms or software packages never posed a problem. Problems with opening or conversing older files are rare, but happen.
- Changes in hardware, software or operating systems is poorly documented
- All firms have some back up policy, using cd-r, dvd, hard disk, tape or an external service
- None of the firms have a policy for refreshing of the data carriers.
- None of the firms have a policy for migration to newer versions or formats. Migration occasionally takes place when presentation documents are produced.
- Regarding authenticity, most firms rely on the fact that data carriers (cd-r) or formats used (pdf) are write protected
- None of the bureau uses archival software
- Storage of project metadata is custom made and not supported by standards

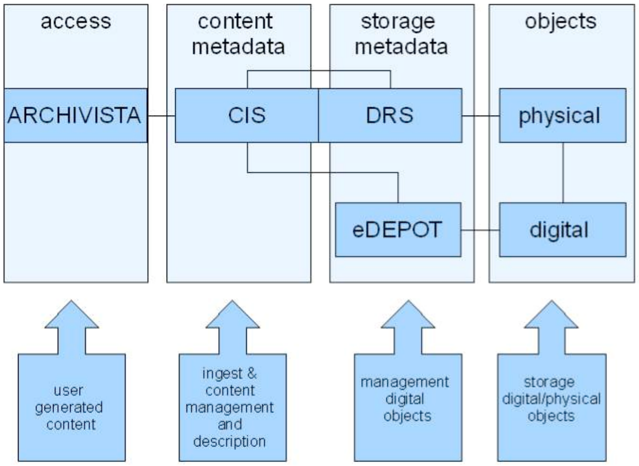
The overall conclusion is that firms in general are aware of the possible problems with sustainability of digital objects, both on the hardware as on the software level. But their interest in this matter is mainly motivated by legal, practical or promotional issues, and less for historical reasons. The effect is that the concern is limited to a medium term (up to 10 years).

The firms were also asked to send a sample of their digital files, in which the documentation of one or more projects was gathered. This gave us an insight in the way firms store, archive and organize their files. Analysis of the files affirmed what we had learned from the interviews. We also learned that:

- Project files hold a lot of copies and derivatives (e.g. low resolution copies)
- Names of folders and the folder structure often offer essential information, supplementing the information provided by filenames and other (rare) metadata.

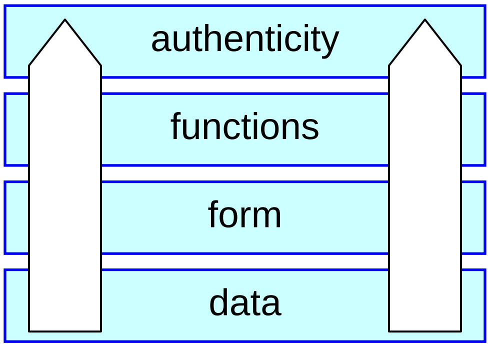
The CIS connection: a two tier approach

Although the Minisis software suite provides some functionality in the storage and description of digital objects, it was clear that the existing CIS couldn't provide all the necessary functions. We therefore decided to develop a independent environment, that would deal with the storage and preservation issues, while the CIS would deal with the content metadata. This allowed us keep the registration of both physical and digital objects in one database, using the same authorities and keyword sets in a straightforward way. In this view, physical, digitized and born digital objects are all treated the same way. The only difference is that analogue objects have a physical storage space which is documented in a location field, while digitized or born digital objects are retrieved by means of an ID or a hyperlink that connects the descriptive record in the CIS with to the record in the repository database. In the Archivista environment, the hyperlink can be used to retrieve the digital object (or a copy).



Towards a preservation strategy

With an ever growing number of digital material in foresight, decisions have to be made about what will be archived, and how this can be achieved in a cost efficient way. The overall aim is to save the maximum of information, which demands for a clear definition of what (relevant) information is. At its most basic level, information is regarded as *data*: text, numbers, forms, sound, etc. A more complicated level is when the form or design in which the information is presented, plays a role in the understanding of the information. Thus it may be important to preserve the formatting (*look and feel*) of a document, or at least its main features. The combination of data and form may be regarded as the equivalent of 'intellectual content' On a third level, a preservation strategy is concerned with the functionality of a document. A three-dimensional CAD object can be viewed in different angels, its shape can be modified, etc. Saving it as a raster based image would mean a significant loss of information. And finally, the authenticity of a document can be questioned, especially when legal issues are involved.



Therefore, a preservation strategy must define for each type of object the desired level of information that has to be retained.

Fundamental to the preservation strategy of the NAI is its central mission "to document the design process." An implication is that the focus is on the data and the look & feel of a document, with relatively less care to functionality or legal issues. On the other hand, the longevity of a document is much more important than it presumably is in an architectural bureau. A preservation strategy should be consistent with archival standards and adapt open formats.

This is complicated by the fact that the NAI is not in a position to impose best practices or archival standards to the firms from which it acquires the archives. Thus the NAI is confronted with a wide variety of formats, whereby different documents are often copies of the same original. This led to the concept of *milestone documents:* in a architectural practice it is common to produce documents at the end of a design phase, often intended as a presentation for a third party audience – and thus often in a more widely accepted format. The consequence is that it may not be necessary to treat every document at the same level of preservation.

These consideration lead to the following fundamentals in our preservation strategy:

- *Open standards*: for every proprietary format, a corresponding *de jure* open format is defined, thus bypassing future technical or juridical limits
- *Migration* as a fundamental preservation strategy, to ensure the longevity of the archive. Priority is given to milestone documents.
- *Delayed migration*: less important documents are migrated on an ad hoc basis.
- *Multiple migration*: since it is not always possible to select an archival format that supports the desired information level, documents may be migrated to more than one format.[7]
- The *original bitstream* of a document is always preserved, thus ensuring future migrations to more suitable/more open formats
- Priority to preservation of *intellectual content* (over functionality and authenticity)
- *Technology watch*: continuous effort to document evolutions in (open) formats.

Building the test environment

The experience from our survey and the analysis of the sample data proved that we needed more hands-on experience with archiving of born digital files. It was therefore decided to set up a test environment that would provide all necessary functions of a trusted digital repository.
A preliminary list of desired functions was based on desk research and analysis of acknowledged standards, such as OAIS[8] and PREMIS[9].

The test environment should allow to:

- Virus check
- Log the folder structure and files on the original data carrier
- Identify and validate the formats of the files
- Check for duplicates
- Migrate proprietary formats to acknowledged archival formats
- Safely store objects and their migrations
- Provide some way of linking between the CIS, ArchiVista and the eDepot itself

Several software packages, most of them open sources or freeware, were tested and evaluated. This resulted in a band-aid solution: the eDepot is running, but not as smooth and efficient as it should. Every action has to be initiated manually, metadata have to be copy-pasted into the proper fields, etc. this provisional set up has allowed us however to define the fundamental questions and principles of the NAI eDepot project.

Virus check

Virus checking is obviously an essential procedure, but is often incorporated in the overall ICT-architecture of an organization. In the case of the NAI, it was sufficient to double check standard settings of the anti-virus software. Since the NAI rarely acquires very new digital archives, it was decided that a quarantine period of three days was sufficient.

Folder structure and file name logging

As a first ingest action, a list was built of the files and the folder structure of each data carrier (in this case, cd-rom). We used a freeware tool called File List Generator, which reported file names, extensions, folder name and path, size and date.[10] This information can later be used when analyzing and describing the content of the archive in the archival module of the CIS.

De-duplicate

As mentioned earlier, the sample files contain a lot of redundant data in the form of copies or derivatives. Sometimes these copies are meaningful and should be kept, as is the case with physical archives; a document can be part of a folder with a selection of a larger collection that is stored in another folder. The location in the folder with selected materials is an important piece of information. But on the other hand, digital objects are often replicated in lower resolution or a different format. Careless archiving often even results in identical files stored in different folders.

Tools to detect these duplicates or semi-duplicates allow to search on different features, such as file name, file date, length or content – the latter comparing each file byte by byte.[11]

Identification and validation

A superficial examination of the sample data showed that objects aren't always what they seem to be: quite often a file extension doesn't match with the actual file format. File extensions don't contain software version information, and some extensions are shared between two or more file formats. Luckily, the open source community has provided several tools identify and validate file formats.[12] Validation is based on the 'binary signature' of a file, i.e. the essential features of a file that are determined by its format. Even when a file extension is missing or wrong, the binary signature allows to determine the exact format and version of a file.

Because of the presence of many proprietary formats, JHOVE proved not to be very useful, since it recognizes only a limited number of formats. DROID, which uses the PRONOM database from the UK National Archives is a better choice for this reason. Around 90% of the files were recognized and validated.

Migration

Once files are identified and validated, the next step is the process of migrating the files to a preservation format. As our draft archiving policy recommends, each file is archived in its original format. It is then migrated to one or more archival formats. The consequence is that the eDepot needs a number of software packages, to allow files of each format to be opened and migrated to another format. Although most files can be opened in other versions of the same software, there's always a possibility that built in conversion tools aren't capable of preserving all information or functionality[13]. Ideally, every version of every software should therefore be available, as to limit the possible data loss.

In reality however, such a software museum approach is too expensive and impractical. The licensing cost would burden the project budget too much, not to speak of the maintenance cost of operating systems and even hardware.

The next best thing to do is to rely on the built-in conversion or migration tools that most software packages offer.[14] A limited number of powerful software packages can be capable of handling most of the files. Especially with 3D vector based formats (CAD files), this approach has some pitfalls. Since the specifications of the native formats are not open, many packages use an intermediary format as a translator. Often these intermediary formats support only a limited number of features. In other cases, the developer has broken the proprietary code – but it is not sure how successful this was. The problem is that the method of conversion is rarely documented by the software vendor, which leaves it up to the digital archivist to test and analyze conversions or migrations.

Creating core metadata

Most of the tools we described above generate logfiles with some essential information on the digital objects. This information should be stored as metadata in the repository and/or in the CIS. Other log files should be kept too, but not necessarily as metadata with the digital object (e.g. virus check log data).

Several standards (OAIS, Premis) provide guidelines as to which of these data is to be stored, but it is up to the organization to decide how and where this is done.

In the process of ingesting digital archives, the separation between content and preservation is retained: as is often the case with physical archives, archival materials are securely stored first, and only after a while the archivists starts assessing, cleaning and describing the archive. When doing so with the digital archive, the logfiles c.q. metadata provide the core information for the archival record in the CIS.

Storage

The choice for a two tier model, where the Collection information system holds all the content metadata, limits the requirements for the eDepot application. Some of the fundamental requirements include:

- Safe storage/preservation capabilities
- Suitable all kinds of formats
- Bitstreams uniquely identified
- Batch import and processing capabilities
- Versioning functions and capture of relationships between bitstreams
- Metadata upload function
- Interoperability: versatile dissemination options
- Extensibility: ability to integrate external tools with the repository

The full list is being written in the functional requirement report.

In the testing environment, DSpace was used, mainly because of the straightforward installation routine and the availability as open source. Although it performed well in our testing environment, DSpace has some drawbacks when it comes to automating tasks. To meet all the requirements, DSpace should have to be rebuilt significantly. We are therefore considering and reviewing other document management systems or digital asset management systems.

The result of this conglomeration of tools is a more or less working model of a digital repository. Although more a toolbox with loose components with little support to workflow processes, this test bed allowed us to develop a clear view on the fundamental questions and practical requirements of a working repository.

Conclusion

On the side of the content creators, the situation is complicated, but not dramatically. Based on a limited survey, we concluded that the awareness of digital sustainability is present with architects and architecture firms. The main threat is the shorter period of time that architects take into account. To ensure the readability of older data (i.e. >1 or 2 decades), the NAi should change its acquisition policy (acquire sooner) and/or try to involve or even educate architectural firms.

The different character of digital archives (as compared to the analogue), implies different ways of appraisal, selection, processing and description of archives. It demands for written policies that determine what archival formats are suitable for a given proprietary format. Such a migration policy framework is highly dependent of the emerging formats and software packages, and therefore constantly in review.

Over the last decade, the development of software tools, standards and best practices has resulted in the availability of all necessary building blocks for the creation and maintenance of digital repositories. However, the availability of out of the box solutions is still rare, especially when it comes to atypical collections with a wide range of proprietary formats, such as architectural archives. The hybrid character of our collections and the cohesion between objects, book and archival material is significantly different from the situation in a typical academic repository.

[1] Library-Archive-Museum, see Zorich, Diane, Günter Waibel and Ricky Erway. 2008. *Beyond the Silos of the LAMs: Collaboration Among Libraries, Archives and Museums.* Report produced by OCLC Programs and Research.
Published online at: www.oclc.org/programs/reports/2008-05.pdf

[2] www.minisisinc.com

[3] E.g. Categories for the Description of Works of Art (CDWA) for the description of objects, ISAD(G) for the description of archival records or ISBD for bibliographic records.

[4] www.mda.org.uk/spectrum.htm

[5] www.nai.nl/cis

[6] Commonly known as Adobe Creatieve Suite, with Photoshop, Illustrator, InDesign, Acrobat, Flash and Fireworks

[7] An example is the .dwg format (AutoCAD's proprietary format), which is not an open standard, but widely supported and offering a high information and functionality level. Migration to .dwg can be supplemented with STEP or SVG.

[8] Reference Model for an Open Archival Information System (OAIS), published online at http://public.ccsds.org/publications/archive/650x0b1.pdf

[9] PREMIS Data Dictionary for Preservation Metadata version 2.0, Published online at http://www.loc.gov/standards/premis/

[10] Download available at http://www.portablefreeware.com/?id=1171

[11] An example is DuplicateFiles Searcher 2.2 (http://duplicatefilessearcher.net), but many more can be found on the web.

[12] Format *identification* is the process of determining the format to which a digital object conforms. Format *validation* is the process of determining the level of compliance of a digital object to the specification for its purported. See the JHOVE website for more on this matter (http://hul.harvard.edu/jhove/)

[13] The term 'conversion' indicates the mutation from a file to a newer version of the format, while 'migration' is used when the file is changed into a different (archival)format)

[14] As a example, OpenOffice 3.0 is able to read MS Word, WordPerfect, StarOffice, of WinWord