RLG DigiNews
RLG's online newsletter for digital imaging and preservation

December 15, 2005, Volume 9, Number 6 ISSN 1093-5371

Building a Digital Archive: A Dutch Experience

Authors: Peter Horsman - Archives School, Amsterdam (p.j.horsman@archiefschool.nl), Klaartje Pompe - Rotterdam (k.pompe@gar.rotterdam.nl)

Note: This is the full text of an article originally published in *RLG DigiNews*. (More publishing information...)

Introduction

In January 2004, the municipal archives of the Dutch harbour city of Rotterdam started an ambitious project to construct a solution for the long term preservation of its digital materials. Project planners asked the Netherlands Archives School for support, and a cooperative effort between the two organizations ensued. The Archives School brings both knowledge and training skills to the project, as well as research expertise. In the School's strategy, teachi research, and practical applications are closely connected.

The municipal archives of Rotterdam is not a small organization. It has a staff of over 100 and holdings of 17 linear kilometers of records, hundreds of thousands of photographs, maps, drawings, prints, and books. Like most Dutch archives, it endorses the Total Archives concept, aiming to document the local society and serving as a primary sourc of the town's rich history, dating back to the Middle Ages. Today, Rotterdam is a modern city, with a population of around 700,000 and a port that boasts to be the world's largest. Increasingly the town administration carries out its business electronically, and consequently, records are created in digital form. The archives anticipates ingesting electronic records in the near future; most of the photographs, television and sound recordings are currently deliver in digital format. Furthermore, like many cultural heritage institutions, the archives undertakes programs to digitize original analog materials in order to improve access and use of its holdings. All of these types of digital materials nee to be preserved as valuable cultural assets.

The initial goal was to develop a low-cost digital repository (e-depot), based on state-of-the art theory, standards, and best practices. The under-lying idea was that if it is impossible to build a digital repository for a reasonable cost then keeping digital materials would be an impossible task for any Dutch archive.

Project Goals

The initial goal was to develop a low-cost digital repository (e-depot) based on state-of-the art theory, standards, and best practices. The underlying idea was that if it is impossible to build a digital reposito for a reasonable cost then keeping digital materials would be an impossible task for any Dutch archive.

During the course of the project the objectives were broadened. The original goal remained the same, but equal importance was given to learning, improving methods, and changing work processes – that is building a competent organization that is able to deal with all kind c problems associated with long preservation and access to digital information.

First Year: Just Do It

For each of the two years that the project has been underway, projec staff chose a different metaphor and associated approach. The first y theme was: "Just Do It," learning by doing. None of the archives staf had any real experience with digital preservation, and the theoretica knowledge was limited. Since the very beginning, all departments ha been involved in the project, including archives management, the ma and print department, the library, preservation, and reference servic The heads of the departments and a project manager formed the project team, and senior staff control the project as a steering committee.

The approach was extremely pragmatic. Based on recommendations from HP Netherlands (the company sponsored the project by provid a server), DSpace was installed. This open source digital library software was already in use by six Dutch university libraries, including the library of Erasmus University in Rotterda The project team attended a DSpace demonstration at Erasmus University, and their digital library staff was helpful answering questions. Also, prior to working with the software, the project team paid a visit to the UK National Archives, to see their system at work. This visit helped shape the discussion about the impact of the initiative on the organization as a whole.

Time Boxing

The project management method the team used during this first year was *Time Boxing*. They divided the work into t fixed periods, or Time Boxes. Each of the Time Boxes had a goal to achieve, but time was the dominant factor for project management. The first Time Box concerned installing, configuring, and becoming familiar with the software. Instead of the standard DSpace user interface, an open source harvester was installed: i-Tor, a Dutch product based o the OAI protocol. DSpace and i-Tor were installed without customization, except to add the municipal archives' logo give the project team the feeling of ownership. The first lesson learned was that the software was not too complicate nothing to be afraid of.

One conclusion from the first Time Box activities was that the standard

DSpace workflow was not suitable for an archive. DSpace expects authors to submit a publication for ingest into a digital library. Such a submission must subsequently be approved by the librarian. The workflow in an archive is different. It is the archivist who decides about capturing records – the ingest strategy is a pull, not a push, strategy. The workflow was seen as too bureaucratic for archives staff with sufficient level of responsibility. Fortunately, a few university libraries felt the same need to change the workflow and the costs of modification could be shared.

One conclusion from the first Time Box activities was that the standard DSpace workflow was not suitable for an archive... Fortunately, a few university libraries felt the same need to change the workflow and the costs of modification could be shared.

The second Time Box continued the testing of ingest and description of a variety of single documents. By the end of the year, the project team was competent to evaluate the DSpace and i-Tor functionality.

Migration

After examining the existing literature on long term preservation, the Steering Committee decided to adopt migration and standardization as major strategies. As much as possible, documents would be migrated to XML. The team acquired and tested an open source migration software utility, XENA, from the National Archives of Australia. The best result was that the team got hands-on experience and insight in the process. A second, less positive result was the finding that the software is still immature.

The project also borrowed much from the hardware architecture of the National Archives of Australia, with separate working environments for quarantine, storage, preservation and description, and access. The fully operational hardware infrastructure was installed in January 2005.

The systems design is heavily based on the Open Archives Information System Reference model, the InterPares Preservation task force model, and emerging metadata schemas.

System Design

In this project, systems design follows development. This is a controversial approach, but it worked well in an experimental, prototyping environment. The systems design is heavily based on the Open Archives Information System Reference model, the InterPares Preservation task force model, and emerging metadata schemas. The overall design approach is based on the Yourdon method of structured systems analysis and design and consists of two series of documents (models) each describing a facet of the system. The first series form the Essential Model or Logical Design. Components of the Essential Model are the Environment Model (describing the interaction between the system and its environment), the Process Model (describing the systems processes), the Metadata Model, the Control Model (describing how the system will be managed), and the Knowledge Model; which is an extension to Yourdon, describing the rules and domain knowledge necessary for running the system. The second series is the Implementation Model or Technical Design. It describes the hardware and software environment, the processors, the software modules, and the human-machine interfaces.

In the first year, a draft of the Essential Model was completed. This model clearly showed that a digital depot was just one building block in the whole of the information architecture for the archives, reflecting the fact that senior management expressed the policy to integrate digital archives management into the normal business of the departments. The project team rejected the alternative strategy of establishing a separate digital archives.

Second Year: Island Hopping

The objective for the second year was to test the limits of DSpace. It was expected that by the end of the year a decision could be made as to whether or not DSpace had sufficient functionality to serve as the basis for the digital depot system.

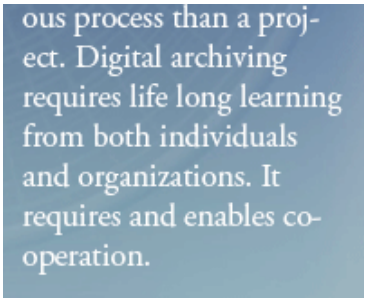
The objective for the second year was to test the limits of DSpace. It was expected that by the end of the year a decision could be made as to whether or not DSpace had sufficient functionality to serve as the basis for the digital depot system.

The project team developed a different project management strategy, using the metaphor of *Island Hopping*. Several research areas were selected, such as bulk data, multimedia, websites, databases, relationships with records creators, hardware and software, policy, access (digital reading room), metadata, procedures, and systems design. A working group was established for each area. User participation in the project this year was increased from 12 to over 30. The team invited other archives to have staff participate in the working groups. Each group had the task to accomplish a predefined goal: find an answer to a research question; that is, "to conquer an island." Some working groups were more successful than others and all had different approaches. For example, some groups tested software, while other groups analysed theories and best practices.

Staff members realized that implementing digital archiving is not just a complex and multi-disciplinary task, but it changes the organization and is more of a continu-

Results

The overall positive result was – I should say "is" – the enforcement of the learning capacity of the organization. Staff members realized that implementing digital archiving is not just a complex and multi-disciplinary task, but it changes the organization and is more of a continuous process than a project. Digital archiving requires life long learning from both individuals and organizations. It requires and enables cooperation. Departments that worked for decades within their own disciplinary lines discovered each other's methods and systems and were open to challenging their own practices and standards. The working group on access provides one example. The group developed functional and



quality requirements for a digital ("virtual") reading room. In the process, the group evaluated existing applications and Web pages and concluded that a redesign would be needed. This appeared to be a serious problem because the different departments not only used different software applications, but variations in metadata models hampered easily understood access to the collections. The group set out to develop a harmonized metadata model for descriptions, applicable to all kinds of materials held in the archives, including records, photographs, prints, drawings, books, magazines, sound recordings, moving images, etc. They started with the ISAD(G) and ISAAR(CPF) standards, endorsing the concept of authority control, a rather unknown concept at that time. The group was successful, and the logical metadata model could even be mapped onto EAD, EAC, and Dublin Core, opening the way for the development of an organization-wide XML meta-database of descriptions, based on open standards, and accessible to harvesters, browsers, and search engines. The implementation of the database should be complete by the end of 2006.

Working with one metadata model affects the applications that the archives is using for description and access. Most of these applications are specific to one type of material and use different metadata schemas. Discontinuing these is not an option, at least not in the short term. Output scripts will likely be needed to migrate the local databases into the generic XML meta-database.

DSpace and Other Software

During the pilot projects carried out by the working groups, the shortcomings of DSpace became apparent. First of all, the metadata schema is insufficient for archival purposes. DSpace supports qualified Dublin Core, which might be good enough for presentation and retrieval of Web publications, but does not meet archival requirements. Full, multilevel description is not possible, nor is authority control. However, DSpace is now available on an Oracle platform, which makes it technically possible for the Rotterdam archive to merge with existing metadata systems. This means intense customization of the software. A second problem with DSpace is the ingest process. DSpace uses an upload procedure that appears to be too slow for large files, such as multimedia data and multi-layered bulk. One of the project's most important questions now is whether or not to continue with DSpace. At the moment, alternative open source software is under consideration. The project team works on a component based architecture for the archival institution as a whole. Much of the required functionality for managing digital archives is not different from managing analog materials, and much of it is already available in current applications. The repository function is just one part of it, and possibly, software is available.

How to Proceed?

Whatever decision about DSpace is taken, the work will go on. Next steps include a rigorous system design. Based on design decisions, software options will be evaluated. The archive will need to decide the extent to which to continue with open source software. The original choice for DSpace and other open source software was a pragmatic one. DSpace was available and proved to be a good tool for acquiring knowledge and experience. A fundamental discussion about open source software has not, as of yet, been undertaken.

Working with open source is not necessarily less costly than using commercial software. It requires working with communities, and there are problems with an archival institution entering into a community mainly consisting of libraries. For open source to work, it is absolutely necessary to establish an archival community to cooperate and to share experiences and costs.

Working with open source is not necessarily less costly than using commercial software. It requires working with communities, and there are problems with an archival institution entering into a community mainly consisting of libraries. For open source to work, it is absolutely necessary to establish an archival community to cooperate and to share experiences and costs.

> Publishing Information

RLG DigiNews (ISSN 1093-5371) was a Web-based newsletter published from April 1997 through April 2007 which was conceived by the RLG preservation community and developed to serve a broad readership around the world. It was produced by staff in the Department of Research, Cornell University Library, in consultation with RLG and was published six times a year at **www.rlg.org**.

The issue of *RLG DigiNews* in which this article originally appeared was produced by: Co-Editors: Anne R. Kenney and Nancy Y. McGovern; Associate Editor: Robin Dale (RLG); FAQ Editor: Richard Entlich; Contributor & Copy Editor: Ellie Buckley; Production: Jenn Colt-Demaree, Carla DeMello; Advisor: Peter Hirtle.

The full issue in which this article originally appeared also is available.

(Return to top...)