FROM STAND-ALONE PRESERVATION TO CROSS-INSTITUTIONAL COLLABORATION

Jan Dalsten Sørensen
Danish National Archives

1. Introduction

The digital revolution where more and more digital content is created, collected and preserved, presents all cultural heritage institutions with similar challenges in terms of digital preservation. Even though the collections in e.g. archives and libraries are different and even though our practices in terms of curation, registration and preservation strategy cannot be identical due to differences in legislation, user requirements etc., the task of preserving the raw bits and bytes of digital material is in most respects the same.

Digital storage at the Danish National Archives has evolved quite a bit over the years. We have used a number of combinations in terms of media types, ranging from large magnetic reel tapes to CD-R's and later DVD's combined with disk based storage. One important characteristic of the storage solutions has been that they have been all defined by the National Archives, and most of the software used in the processes of test and transfer to the storage media was also developed in-house.

A number of years ago, however, the Danish National Archives began to outsource the bit preservation of one of its copies of digital records to the State and University Library in Aarhus, Denmark. The basic idea was that shared infrastructure for preservation would make it easier to achieve economies of scale and would lead to a more cost-effective digital preservation for the cultural heritage institutions involved.

In line with this, the key institutions in the field of digital preservation in Denmark - the National Archives, the Royal Library and the State and University Library - decided to explore the possibilities for further cross-institutional cooperation in the field of bit preservation.

In an exploratory project, the so-called Bitrepository-project, the three institutions worked together to specify the requirements for a shared architecture for long-term preservation that would let them share preservation facilities while still maintaining full control of their own collections in terms of access, intellectual rights, logical preservation etc.

This article will sum-up the findings of the project and describe the subsequent implementation of the Bitrepository at the Danish National Archives.

2. The Bitrepository-project

The partners in the Bitrepository-project have similar tasks in preserving the digital cultural heritage. However, their collections and requirements are different in a number of ways.

The National Archives collects and preserves born-digital records from the national, regional and in some cases local authorities. We also digitize a growing number of paper records from our collection so they can be made publically available on our website through the service "Arkivalier Online" ("RecordsOnline"). The Royal Library and the State and University Library collect and preserve all sorts of digital publications as well as radio and TV-broadcasts. The libraries are also responsible for the harvesting and preservation of the Danish part of the Internet. The libraries are, of course, also digitizing large parts of their holdings. Thus, the collections are very diverse in terms of size, provenance, restrictions for use, uniqueness, demand for access etc., etc.

The Bitrepository-project needed to define the necessary levels of service and security that the partners needed for their different collections. The project also looked into what kind of architecture and organization would support the identified needs.

In order to define the most suitable architecture and technological set-up, the developments in the markets for storage solutions and similar projects with shared infrastructure were explored. For instance, the project looked at, and was very inspired by, peer-to-peer projects such as LOCKSS ("Lots of Copies Keep Stuff Safe") and CLOCKSS, the controlled version of LOCKSS for the shared preservation of digital information that is not publically accessible.

For preservation institutions, bit preservation solutions must include support of bit safety as well as other requirements like e.g. confidentiality and availability. It must be possible to adapt these requirements so that they suit the collection in question. Thus, the architecture of the Bitrepository should support all sorts of combinations so as to ensure that the various requirements could be met for each collection in this "one-size-does-not-fit-all"-environment.

The architecture

The Bitrepository-project came up with an architecture where each institution operates one or more ”pillars” where the digital material is stored. The pillars should be completely independent and based on different preservation technologies and media. The pillars considered in the Bitrepository-project included preservation on tapes, on DVD's, on Direct-attached Storage (DAS) and in a SAN (Storage Area Network), but the architecture is so flexible that other types of pillars, including a cloud pillar could easily be integrated. In the proposed architecture, data is replicated between a number of pillars in as many identical copies as the collection owner finds necessary.

Each institution has its own ”client” which exchanges information with the pillars of the Bitrepository though a thin so-called ”middle layer”.

The key to make sure that the bits remain unaltered is to keep checksums of all copies of data and to perform checksum checks across the pillars. Thus each institution can continuously keep track of the checksums of all copies and react if unintended alterations are detected.

The set-up with independent pillars, built on different technologies, owned and operated by different organizations in different locations counters the most obvious threats for bit preservation. In this set-up, the preservation of the digital material is not dependent on one particular technology, one particular commercial product or vendor or even one organization. The idea is to avoid any single point of failure. A natural disaster or any other disaster, like for instance a fire, will not destroy all copies of your digital material. The involvement of several organizations makes it impossible for one person (malevolently or accidentally) to affect all copies of a particular file in the Bitrepository.

Logical preservation

The idea of the Bitrepository is to provide the institutions with bit preservation, but it does not guarantee that the bits can be interpreted! It is still the responsibility of the collection owners to make sure that adequate metadata about the content of the collections is created and kept; to make sure that the digital material is migrated to new formats when necessary etc.

Each pillar-owner will negotiate Service Level Agreements (SLA's) with its customers. The SLA should in principle define all the requirements of the collection owner from response time to security and audit trails.

Recommendations

The Bitrepository-project recommended the three partner-institutions to devlop a shared infrastructure for bit preservation, built on the above-mentioned architecture. In the actual development and implementation of the Bitrepository, several important points should be taken into consideration, such as:

- Avoid “single points of failure”, understood as critical components that perform vital functions in a system alone. If a component like that fails, the whole system fails.

- Use of different types of software, hardware and media to avoid systematical failures based on shortcomings of particular types of technology

- No one person should be able to affect all copies of a particular file in the Bitrepository

- Several copies are, of course, necessary. The number of copies will vary depending on the nature of the digital material in question.

3. The implementation

Based on the results of the initial project, the three institutions decided to go ahead and develop software and implement the proposed infrastructure. The software should be open source so as to be able to eventually attract more project partners that would be able to help develop and maintain the code.

The Bitrepository software was developed by systems developers in the three partner institutions. One institution used c# while the others used java, again in order to avoid any single point of failure. The Ministry of Culture granted approximately 525.000 € to the project. This sum has been used almost exclusively to pay the staff involved in the project. The further implementation and development will have to be defrayed by the institutions themselves. The investments in the necessary hardware have not part of the Bitrepository project as such. Hardware for storage has been purchased by the pillar owners and the cost will of course affect the price for storage in each pillar. Each institution has to pay for storage in the other institutions' pillars.

The purpose of the Bitrepository system as it has been implemented over the past couple of years is “to enable long-term preservation of data in a distributed, highly redundant architecture. The data integrity is ensured by using multiple, independently developed data storage systems (pillars) across different organizations, together with functionality for maintaining the integrity of the data over time”[i].

The Bitrepository is first and foremost to be used by the three institutions behind it, but it is also developed so flexibly that it will be able to provide bit preservation services to other interested institutions.

In the following, this article will present the digital collections at the National Archives and describe how the collections will be preserved within the Bitrepository framework.

The digital collection at the Danish National Archives

The digital collection at the Danish National Archives can primarily be divided into two parts: the born-digital records and the digitized records.

The Danish Government has on all levels become almost exclusively digital. The National Archives is the institution that receives and preserves the records from the government, or rather, all of the national government and about half of the municipalities. The digital records from the authorities make up the born-digital collection. The second part consists of all the digitized records we have. This is by far the largest part of the digital collection, as it has been an important focus for us, like for so many other archives, to make to most frequently used records available to the public on the Internet.

The amount of born-digital records is currently about 55 TB. At the beginning of 2014, we had about 40 TB, and we expect that the accession this year will total about 27 TB. This is about three times as much we have ever received and processed in any previous year and shows that it is necessary to continually expand and develop our preservations systems to cope with the growing amounts of data. The amount of digitized records is about 100 TB, and this collection is also growing rapidly.

Seen from a pure bit preservation point of view, you can argue that it is unimportant whether your digital material is born-digital or digitized. However, the necessary levels of security, access etc. may vary a lot, and as an institution with limited budgets you simply cannot afford to pay for better service and security than what you actually need.

To determine what level of security and service is necessary for a particular collection, it is a good idea to take some of the following parameters into account:

- Confidentiality: What will happen if an unauthorized person either by accident or intentionally gets access to the information?

- Accessibility: How often do you need to access the material, and what is your level of tolerance in terms of response time?

- Integrity: Are the records unique or are they digital copies of material that could be re-digitized?

In terms of confidentiality, there is a big difference between our born-digital collection and the digitized records. The majority of our digitized records are already publically accessible on the internet, so here there are no specific considerations about confidentiality or secrecy. When it comes to the born-digital records, however, the situation is quite the opposite. The born-digital records that consist of data and digital documents that have been submitted from the it-systems of the government are not accessible until after 20 years or, if they contain personal data, after 75 years. So, while the majority of our digitized collection is openly accessible, the majority of our born-digital records must be treated as confidential. This is of course an important parameter for the necessary level of security for the two collections.

When it comes to accessibility, i.e. how fast and easily you should be able to gain access to your stored data and retrieve them from the repository, we can usually work with pretty long deadlines. We hardly ever have to be able to retrieve data within minutes, so we can rely on near-line or off-line solutions which are much less expensive than disk-solutions.

Also when it comes to integrity, there is a gap between the born-digital records and the digitized records. The digitized records are just extra copies of original parish registers, census sheets etc. In the worst case scenario, where the digital copies are lost, you will be able to redigitize the material. You will of course lose the original investment made in the digitization of the material, but the integrity of the records is not at risk as such. Unless, of course they are so fragile and affected by mold or similar physical conditions, that the paper copy is in the process of getting useless. In that case the digitized records have to be treated the same way as the born-digital records. If everything goes wrong with your born-digital records, they are lost forever, unless they by chance are still kept by the records creator. This makes it important for us to set higher demands for the protection of the integrity of the born-digital records.

When you look at parameters such as confidentiality, accessibility and integrity, you need to realize that your demands and wishes must be seen in relation to the costs. You can always wish for a lot of things, but you should also be willing and able to pay for it. Since the money that we spend on digital preservation comes from the tax payers, we need to define our requirements carefully and make sure we do not spend more money than necessary. Why should we pay for an expensive on-line disk-based storage solution with easy and fast access for digital records that are not frequently used and where there is no real need to be able to retrieve the data within a very short time span? Likewise, why should we have as many copies of digitized material that is not unique as born-digital records? The costs of storage should certainly not exceed the costs of a re-digitization!

The implementation of the Bitrepository at the Danish National Archives

For the digitized records the general strategy is to have one copy on disk in our own SAN and one off-line copy on tape in a pillar that is hosted by the State and University Library. For pure preservation purposes you might argue that it would be sufficient to keep the preservation copies on near-line or off-line tape but that would go against the principle of using more than one type of technology.

For the born-digital records the strategy is to have two off-line copies on DVD and one copy in a near-line tape pillar hosted by the State and University Library. It is, of course, no secret that the DVD medium is a challenge for us. First, because some of the submissions that we receive now are so big, that we literally need thousands of DVDs to store the information. This of course implies a lot of manual handling which is expensive. Also, the DVD may begin to lose its importance as storage medium which means that there are less different brands and products to choose from, which again increases the risks of us becoming dependent on one producer, something that we work very hard not to.

The reason why we so far continue to use DVDs is mainly because of our principle of risk spreading: We would really like to use both optical and magnetic media in our preservation system. If we use only magnetic media it will imply some risks that we are not keen on taking. We might use BD (Blu-ray) instead of the regular DVDs which will reduce the need for manual handling simply because one BD contains more data than a DVD. We are of course also very interested in the prospects of new high capacity archive discs. As an archive, however, we cannot just jump on any bandwagon. We need to be sure that the technology is sufficiently supported and will provide us with a sufficient number of quality brands to choose from.

Cross-institutional collaboration

In the process of implementing a new preservation system for our digital collections in the Bitrepository, we have seen it as a clear benefit that we have been able to have a closer cooperation between the National Archives and the libraries. We now know each other a lot better than before and have a better platform for the sharing of knowledge. Relatively speaking we are all small institutions and we can only benefit from the exchange of knowledge and ideas that happens when you go together in a common project and sharing a common infrastructure for preservation.

However, when three smaller institutions go together in a shared project, there is always a risk of being more dependent on resources in the other institutions. If the IT-department e.g. in one of the institutions lacks available resources due to vacation, leaves of absence or any other reason, it does not only affect that particular institution, it might as well affect the two other partners.

Another lesson is that the actual implementation can be difficult. The challenges of making software and hardware work together across institutional boundaries should not be underestimated!

4. The way ahead

The Bitrepository software is available as open source, and if the quality of the software should be maintained over time, it would be great to have some more memory institutions on board. You can read more about the Bitrepository software and the open source project at www.bitrepository.org.

The consortium behind the Bitrepository is represented in a group called ”Distributed Digital Preservation” through a staff member from the Royal Library. This group consists of a number of influential institutions and groups in the field of digital preservation such as the Internet Archive, LOCKKS, DuraCloud and Library of Congress to name a few. The group has agreed on the following definition of Distributed Digital Preservation:

” Distributed Digital Preservation is the use of replication, independence, and coordination to ensure digital content remains accessible by addressing the known threats through time”

Once it is fully implemented, the Bitrepository in Denmark will be an example of distributed digital preservation that complies with this definition. We hope that our experiences will be of use to the digital preservation community and maybe help prevent too many ”reinventions of the wheel”.

Notes

[i] Bitrepository.org