Mass CD/DVD Migration: A Novartis Case Study

D. Bienz, R. Gschwind, Imaging and Media Lab, University Basel, Switzerland; M. F. Pozza, L. Gantner, Novartis Pharma AG, Switzerland

Abstract

Nowadays computers are an essential part of modern pharmaceutical research and data storage. It is also well known that experimental and observational data is crucial for scientific research. For approximately twelve years now, the pharmaceutical industry is storing its research data on portable storage media, where each research group was obligated to archive their research data on such portable media but they had also much individual responsibility for the archiving process. So far there are over 3.000 portable media in the raw data archives of Novartis. Digital data archives are more unstable than paper-based archives basically due to development of computer technology and rapid changes of digital storage media. To guarantee the future readability and to preserve the research data, Novartis has decided to migrate the data from the current portable storage media to a storage type more suitable for long-term preservation while preserving the authenticity and provenance. In this paper, we would like to report on this migration project that nicely illustrates several problems of digital archiving and provides state-of-the-art solutions for them.

Introduction

Companies in the pharmaceutical industry like Novartis have historically documented their experimental and scientific data in paper-based laboratory notebooks. These notebooks have been afterwards stored in the companies archives for safekeeping. These days computers are an essential part of modern pharmaceutical research and data storage so that today more and more research material is "born digital" [1]. Accordingly, the change from paper-based laboratory notebooks to suitable digital formats has taken place. For about twelve years now, the pharmaceutical industry is storing its research data on portable storage media such as CDs, DVDs, ZIP drives and other kinds of media. Up until now, every research group is obligated to provide all of their research data on portable storage media for the electronic archive but how this had to be done was handled individually. Therefore each research group have been provided with a lot of personal responsibility how to manage the research data.

Consequently the raw data archives were and are still growing instantly and rapidly. Up to the present there are over 3000 portable storage media in the raw data archives of Novartis and still growing daily. It is obvious that the era of computers has converted the way of creating, handling, accessing and archiving experimental and scientific data. To guarantee the future readability and to preserve the research data, the data from the current portable storage media needs to be migrated to a storage type more suitable for long-term preservation while preserving the authenticity and provenance.

The challenges now lie in the rethinking of the archiving process, while at the same time manage the digital data that has to be preserved. The digital formats in use will inevitably become obsolete with time and need to be migrated. The archiving processes that were created for paper based journals need to be adapted to digital data, and not just taken over. Therefore there is a strong demand for a long-term digital archiving solution. In the following, we would like to report on the Novartis migration project that nicely demonstrates multiple problems of digital archiving and presents a state-of-the-art solution.

The remainder of this paper is organized as follows. The next Section discusses the general digital preservation theory and positions of our work inside the preservation process. In Section *Migration Description*, we describe and discuss the migration process as we developed it for Novartis and the problems we encountered during our work, and the last Section concludes the paper.

Digital Preservation Theory

We can describe the process of digital preservation as a communication process with the future ([2] and [3]). This process of communication with the future transmits our digital data over four layers as depicted in Figure 1. Beginning with the information that we want to preserve down to the physical medium onto which this information is stored. To be able to preserve information over time, i.e. the digital data that we want to archive, we have three layers that we need to attend to.

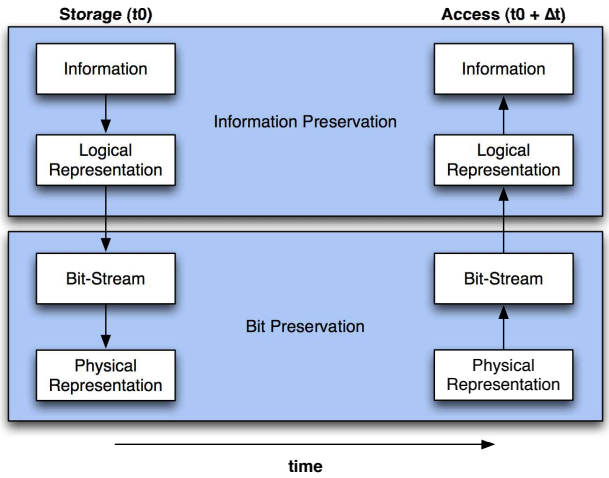

Figure 1. *Digital Preservation Process*

At the lowest level, the physical representation of an archived digital object needs to be preserved over time. Preservation at this level is challenged by the basic decay of the physical media used for storage. For example the life expectancy for optical disks lies between 2 and 50 years (depending on quality and storage

conditions), those of magnetic tapes between 2 and 30 years and for hard-drives between 5 and 10 years.

At the bit-stream level, we have technological obsolescence of the hardware and software used to store the digital objects and which is also needed to access them. This time-span lies between 18 months [4] and 5 years [5].

At the logical representation level we need to preserve the syntactic and semantic rules needed to interpret the bit-stream, without which the digital data stays by itself meaningless. By the same token, digital media storing bit-streams are meaningless to the naked human eye. Without meaning data cannot provide information. In order to become understandable to humans, digital data needs to be interpreted and represented by a computer system. Therefore not only the data itself and the data carrier need to be preserved to guarantee future readability, but also the description for its interpretation and rendering by a computer system.

Digital Preservation of Scientific Data

These days it is common that the information technology is an indispensable part of the research. The documented research data on paper is being replaced by computer-generated digital data [5]. Information technology has changed the scientific research in such a way that the opportunities of how to document data has been expanded. The central issue resides in the fragility of digital data that is the hardware and software become obsolete very fast. The requirements for archiving of scientific data are changing simultaneously with the developments in the information technology. Additionally there always remains the question which scientific data is fundamental for archiving and must be preserved for the future.

For example according to the Novartis guidelines, the authenticity of research data is ensured by having the responsible group manually sign the medium they provide for the archive. Also every research group is instructed to hand over all of their research data on portable storage media for the electronic archive but how this should be done was not specified and ultimately done very individually.

When archiving research data the subject of authenticity and provenance is very important. In the days of the paper-based laboratory notebooks, the researchers where obligated to sign their notebooks and hand them over to the archive for safekeeping. Nowadays with digital "notebooks" it is not nearly enough to just just copy the data to some portable medium and to take over the procedure like in the "old days" and sign the CDs in order to guarantee authenticity and provenance, as we have described in the example of Novartis above. One can take a thirty year old laboratory notebook and flip through its pages to find the needed information, but to preserve digital data for the future including findability and readability, more needs to be done.

Positioning of Our Work

Guided by the narrow scope of the project, the work we are going to describe in the next Section, is situated in Figure 1 in the lower rectangle titled *Bit Preservation*. By migrating away from portable storage media we are able to preserve the bits. This is just a first step in the right direction towards a solutions, since the *Information Preservation* part is still not solved. The final solution needs to attend to all layers like for example in [6] or [7].

Migration Description

To preserve and rescue as many data as possible we have constructed a custom-built script. In this Section we will describe the steps of the migration procedure. First we will illustrate why we used the ISO image format for the data migration and which the key benefits of the ISO image are. Secondly we elucidate briefly the UNIX program `''diff''` before the miscellaneous errors which occurred during the migration procedure will be described in detail.

Description of Data Migration Workflow

For ensuring a proper data migration of the CD onto the hard disk there has to be provided an appropriate format. For this we used customized scripts. At the end we took a picture of the portable storage medium for guaranteeing the authenticity of the particular CD, DVD or ZIP drive.

For each medium we created a small record as follows:

- a hybrid ISO-image
- log file (how the copying process of the data comparison took place)
- entire listing of the content
- picture of the CD, DVD or ZIP drive
- checksum of the picture and the ISO-image

The guidelines of Novartis specify that authenticity of the portable media is given if it was signed by the respective research group. In order to transfer this as well, we photographed every single medium and attached it to the corresponding ISO file, including metadata with the listing, a log-file of the copying process and the checksums.

Given the broad variety of miscellaneous formats for the copy of the portable storage medium finally we decided to use the ISO image. In various literature the ISO image is recommended. There exists some alternatives like the TAR file or even the ZIP file. We wanted a format that runs on all available systems.

Hybrid ISO Image

The disc image of an optical file is also referred to as an ISO image. The ISO image is a format which is defined by the International Organization for Standardization (ISO). The aim of the International Organization for Standardization is to provide important standards for companies and it consists of different national standard organizations. An ISO image is the method how a CD is been described and it usually includes the complete image of a an optical disc. The development took place for many years. The first original ISO 9660 has many restraints, for example the length of the name is regulated. It is not possible to give the ISO 9660 a name that exceeds a certain amount of letters. The ISO standard has been extended via additions (Joliet, Rock Ridge) to overcome metadata and naming constraints associated with the original specification. To be fully compatible with all common operating systems we used the hybrid ISO image to store the data as hybrid ISO images [8] [9].

The advantages of an ISO image compared to other formats are obvious:

- an ISO image is mountable and serves as sort of a "virtual optical disc"
- ISO images are platform independent

The ISO image of the portable storage medium got an archive identifier due to the fact that the name of the medium was not identical with the archive identifier. To clarify these circumstances here is an example with an optical disc: The original volume identifier of the CD is “Backup” or any ambiguous name and the archive identifier of the CD is “example.01”. To ensure an unique identification of the CD the name of the ISO image is “example.01.iso”.

UNIX Program ‘‘diff’’

For the data comparison we used the UNIX program ‘‘diff’’ to compare the original CD with the mounted ISO image. This allows a bit-wise comparison. In case of not existent errors the checksum has been calculated and the portable storage medium has been migrated correctly. By the use of the checksum the image identity can be verified.

Miscellaneous Errors in the Migration Procedure

Our findings show that only about 90% of the portable storage media could be migrated properly and accordingly ran through the standard migration procedure, i.e. the data comparison with ‘‘diff’’ gave identity. This finding is very surprising due to the studies of Youket and Olson [10] that after eight to nine years about 1% of the CDs reached end of life and were not readable anymore in contrast of the approximately 10% unreadable CDs we have found in our project.

There are different factors that can influence the life expectancy of a portable storage medium, such as type, quality of the medium after the production and before burning, inaccurate burning of the medium, storage and environmental conditions [11]. It must be pointed out that some of these elements can be prevented by a proper supervision of the burning process. Some losses may not have a huge influence to the readability of the medium if we ignore damage through gross negligence. As long as the error correction coding can correct all the errors a portable storage medium remains readable. The issue is that it is unpredictable to determine when the life expectancy of the medium ends. The technical advances proceed with a tremendous pace and do not simplify the problems of digital preservation [12]. In fact data migration is needed to be up-to-date with the developments and innovations of the hardware as well as of the software. The disregard of appropriate data migration can cause extensive costs to recover the data. In the following sections we describe the errors we were confronted with within our project.

These errors have different reasons and can be divided as follows.

Blank Portable Storage Media

A blank CD/DVD for example was the consequence of the inexistent supervision and control of the burning process. The indispensable supervision was missing to check if there is data on the CD/DVD. The CD/DVD has been handed over from the responsible of the research group to the archiver and it has been signed by both of them. But the content of the CD/DVD was never tested. We found out that about 0.5% of about 3400 portable storage media were just blank. These blank CDs/DVDs could be easily avoided through proper monitoring of the delivered portable storage media.

Figure 2. *Spots on CD*

Logical Errors During Data Comparison

Mostly during the data comparison process logical errors of various nature emerged. These errors occurred primarily on the software level. For example:

- sessions has not been closed (open sessions)
- recursive links
- links which showed onto the server
- redundant files but with case insensitivity (upper and lower case) or with special characters
- CDs with multiple sessions which required special approach (multi session CDs)

All these incidents caused that the content in fact was identical but the blank link was not there anymore on the hybrid ISO image. About 1 1/2% of the portable storage media had logical errors. Making a hybrid ISO image is done by using a specific program, e.g. ‘‘mkisofs’’ or the Mac OSX system program ‘‘hdiutil’’, i.e. it is a format conversion and in this case these errors will show a difference between the ISO image and the original CD.

Poor Quality of the Optical Media

We found also some physically damaged portable storage media, basically an advancing deterioration of the media. For instance the CD had spots on the surface as in Figure 2 or showed even progressive degradation. Poor quality of the portable storage media was rare, about 0.5%. Never the last most of these CDs were still readable. We only had a problem with one CD with layer decomposition which made it unreadable. As also shown in [10].

Poor Quality During Writing

Most hardware read errors occur during writing or burning a CD. Mostly due to the missing consistency between the CD writer and the blank CD or that the blank CD was written too fast [13]. The result is that particular sectors of the CD can not be read. This is severe if it occurs at the beginning of the CD, then the whole content of the CD could be useless.

From a production run of approximately 530 CDs, about 230 CDs were unreadable. What happened? The responsible of one research group has burned all the CDs (including a second copy of each CD) with a defect CD writer. Unfortunately this particular CD writer has been adjusted inaccurately, i.e. the CDs could not be read on any other drives. Therefore we searched for the original computer with this specific CD writer and hoped to be able to

read these 230 CDs on this machine. But we were only to some extend successful.

In all these cases of errors the use of miscellaneous drives and operating systems was our course of action. Additionally we had to make use of special data recovery software in case of open session CDs.

Conclusion and Further Work

As depicted in the digital preservation theory, the technological obsolescence of hardware as well as software at the bit-stream level requires continual migration of the data to preserve it for the future. The restructuring to an electronic archive as a replacement of the paper-based archive led to some miscellaneous errors as we described, e.g. blank portable storage media, logical errors, poor quality of the optical media, and during writing. Furthermore the analysis of the file types on the portable storage media is an important topic which is still work in progress. For example we could migrate a specific format over and over again, but if the software does not exist anymore a continuous migration is pointless.

The quality of the medium itself plays a relevant role in the successful and permanent storage of research data. As we found out during our data migration process, poor quality of the portable storage media caused some errors. Although good quality can not guarantee a longer life expectancy of an optical disc, at least it can reduce or even prevent progressive degradation or layer decomposition of the optical media. Fortunately there were only a few CDs with degradation in this project. Overall there were around 10% of the portable storage media which caused problems and were unreadable compared to the results of Youket and Olson [10]. In their studies, roughly 1% of the optical discs reached end of life after eight to nine years.

To sum up, we can say that monitoring the archiving process is indispensable for long-term archiving. The electronic part of the archive has to be taken under advisement. Moreover an employee training and a supportive attendance during the archiving process could prevent many errors. The question is also if a long-term archive could exist if it is only based on optical discs. As the research presented that the life expectancy of optical media is not predictable, maybe a different approach of long-term archiving should be taken into account. All things considered there is still a lot of work ahead of us to adapt to the rapid technological changes and to benefit from the technological innovations for long-term archiving.

References

[1] B. Smith, Preserving Tomorrow's Memory: Preserving Digital Content for Future generations, Information Services & Use, 22, 133-139 (2002).

[2] R. Moore, Towards a theory of digital preservation, The International Journal of Digital Curation, 3, 63-75, 2008.

[3] M. Mois, C.-P. Klas, and M. Hemmje, Digital preservation as communication with the future, 16th International Conference on Digital Signal Processing, 1-8 (2009).

[4] T. Kuny, A Digital Dark Ages? Challenges in the Preservation of Electronic Information,IFLA Council and General Conference, 1997.

[5] J. Rothenberg, Ensuring the Longevity of Digital Information, http://clir.org/programs/otheractiv/ensuring.pdf

[6] I. Subotic, S. Margulies, and L. Rosenthaler, DISTARNET: Distributed Archiving Network, in Proceedings of Archiving 2006, 3, 131-134 (2006).

[7] Peviar, http://www.peviar.ch

[8] Wikipedia, http://en.wikipedia.org/wiki/ISO_9660

[9] Wikipedia, http://en.wikipedia.org/wiki/Hybrid_CD

[10] M. Youket and N. Olson, Compact Disc Service Life Studies by the Library of Congress, in Proceedings of Archiving 2007.

[11] F. R. Byers, Care and Handling of CDs and DVDs - A Guide for Librarians and Archivists, NIST Special Publication, 1-50 (2003).

[12] S. Chen, The Paradox of Digital Preservation, Computer, pp. 2-6 (2001).

[13] H. Bennet, Understanding CD-R & CD-RW, Optical Storage Technology Association, 2003.

Author Biography

Daniela Bienz received her Master's degree in Social Psychology, Consumer Behavior and Organizational Behavior from the University of Basel in 2009. She works in the Imaging and Media Lab at the University Basel where she is responsible for the Novartis data migration project.

Rudolph Gschwind works on imaging technology and photography at the Imaging and Media Lab at the University of Basel. Prof. Gschwinds research topics are image processing and analysis, color photography, color imaging, digital archiving, and preservation of the photographic cultural heritage.

M.F. Pozza, Head of Planning & Operations, Research Operations Switzerland, is a neurobiologist by training and nowadays also responsible for archiving processes in the Novartis Institutes for Biomedical Research Site Basel.

Ludwig Gantner is working in the domain of Records Management for Novartis Pharma AG. He holds a master in History, Social Science and IT.