3

Migration

As described in Chap. 1, migration keeps documents accessible and up to date by continuously migrating them onto some current computer environments. Using an example from our daily routine, Sect. 3.1 illustrates different options for the migration process, gives a succinct definition for the migration process itself, and outlines its overall goals. Section 3.2 deals with applications of the migration approach for long-term preservation of digital documents and discusses possible variants. Section 3.3 addresses organizational aspects within the framework introduced in Chap. 2. Section 3.4 summarizes strengths and weaknesses of the migration approach as a whole.

3.1 Migration: Definition and Goals

migration everywhere

In order to access old or archived documents with upgraded software or on a new computer, most experienced computer users have already migrated tons of documents. This is not the only reason for migration: the Internet connects an enormous variety of computer generations of different makes, each with their own individual hard- and software configurations. Whenever Internet users want to exchange documents, they have to find a means – most often by applying migration – to make the documents readable for their partners. Some readers who are used to exchanging MS Windows documents on a daily routine with their business partners may find this idea strange. Within any mono culture like Intel-Microsoft-Office (or, similarly, Linux-StarOffice) sharing documents is easy. In a more heterogeneous environment, however, migration will probably be the best way to make documents accessible to each other.[1]

[1] Even when running the *same* office product on *different* operating systems, say StarOffice on Windows and Linux, or Microsoft Office on Windows and Macintosh, file sharing may be hampered, e.g., by font problems.

for example, text documents

Let us take a look at a typical heterogeneous setting: On one side, MS Windows users producing their documents using Microsoft Word, on the other side, Unix users generating their documents with LaTeX. In this situation, a pragmatic approach would be to migrate the documents to be exchanged into a third format available to both parties like, say, a printer format such as PostScript or PDF. If "exchange" just means to be able to *read* each others' documents (using viewer software or after printing the documents on paper), this simple approach is sufficient. Closer forms of collaboration require that both parties be able to edit the same documents in turn in order to reach at a final consensus. To adopt the format used by the other party (either in the Windows or in the Unix world) would force users to install a new software (possibly on a new machine), to learn its idiosyncrasies, and, in general, give up the work environment they have grown accustomed to. In many cases this is considered unacceptable.

ASCII as a lingua franca

Users, therefore, tend to simply export the textual contents of documents into ASCII representations and, similarly, to import such ASCII representations into their own systems. ASCII is the lingua franca in the world of bits and bytes. In spite of this, important elements of text file representation are encoded differently by different operating systems, thus hampering transformation by introducing subtle errors: In MacOS, Unix, and MS Windows/DOS "newlines" are represented by either carriage returns, line feeds, or combinations of both. Another obvious disadvantage of this kind of migration is that all formatting information such as font type, color, size, etc. is lost and has to be replenished by the receiving site – a time-consuming and error-prone task.

no general migration strategy

Our example shows that migration can be handled in different ways: there are different target formats (in the example PDF and ASCII) with distinct characteristics. Migration steps are performed using a variety of tools on different levels of granularity: single text files can be transformed using export and import filters of appropriate text processing software. Large collections of text files can be migrated automatically between operating systems using suitable conversion software. Due to the large amount and ever growing number of formats for text documents, image files, sound samples, video streams, databases, and due to the growing number of compound documents, i.e., documents created using a mix of the mentioned representations, a single general-purpose migration strategy for all kinds of documents is not conceivable. For the same reasons the term *migration* does not denote one single approach to migrating documents. Instead, the term stands for a whole class of migration strategies, all with slightly different semantics. In the following, we give a succinct definition and differentiate between several variants of the term migration.

definition of "migration"

An often cited definition of migration was given in the final report of the *Task Force on Archiving of Digital Information* (TFADI 1996):

Definition: *Migration* is the periodic transfer of digital materials from one hardware/software configuration to another or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.

Migration is a recurring task!

One important aspect of the TFADI definition is that migration requires periodic transfer, or, to make it even more explicit, migration is a recurrent, never-ending task, not just a single ad hoc transformation step! This results in a number of issues to be addressed by institutions that employ a migration strategy for archiving digital information over long periods of time. Such institutions have to establish organizational structures that allow for (a) regular inspection of the digital materials, (b) planning of the next migration step as soon as inspection detects a migration need, and (c) performing individual migration processes according to plan as part of their daily routine. For archives, libraries and all "data-housing units" within private organizations, migration is a permanent change process. Section 3.3 focuses on organizational aspects of this change process.

broad and strict senses of migration

The TFADI definition describes "migration in its broadest sense," and, thus, identifies the overall goals of any long-term preservation strategy: Digital materials must be kept accessible in an as authentic as possible form for potential future users. As discussed in Chap. 1 in the context of the hardware museum approach, digital documents can be accessed using their original rendition system only for a rather limited period of time. Therefore, either the original documents have to be transformed into formats available on current platforms, or a replica of the original rendition system has to be recreated. Both strategies conform to "migration in the broadest sense." The first strategy is called "migration in a stricter sense," and is the focus of the remainder of this chapter. The second strategy is called "emulation" and will be discussed in Chap. 4.

requirements for migration

Before discussing the different migration approaches used for long-term preservation, we briefly recapitulate the objectives to be achieved. Lievesley (1995) lists the following objectives for the preservation of digital material:

- Physical data reliability, i.e., preservation of the data media without loss of information
- Protection of the data against unauthorized access
- Continuous access to the data and
- Integration into an information brokerage environment

planning for migration

Planning a migration step is difficult, since requirements on future data use in general cannot be anticipated completely. The following questions arise whenever migration is about to be performed: How much

loss of formatting information can be tolerated? What amount of reformatting is acceptable in the current setting? What will be the total cost of the planned migration step? Every single migration step jeopardizes the integrity of the digital material (TASI 2000). Integrity constraints relate to both the original intellectual content and the form and structure of its representation (for instance, resolution and color representation of images). In any case, users of archives need to be certain about the authenticity and the high quality of the preserved digital data. For data transmissions over the Internet, a variety of nonreversible compression techniques (resulting in loss of data) are applied to images and videos in order to reduce the required bandwidth. These techniques are not acceptable for migration as discussed here (TASI 2000). Obviously, from the archives' and libraries' perspectives, preservation costs have to be kept as low as possible. An important question in this context is which data need to be preserved for the future. If under financial pressure we decide to preserve only surrogates of original documents (e.g., summaries), this decision cannot be revised in future.

keeping documents accessible

Michelson and Rothenberg (1992) assume that digital libraries are mainly used for teaching and research. Therefore, it is important that digital documents can be retrieved, analyzed, and further processed at any given time (cf. (Endres and Fellner 2000)). Digital documents must be preserved in a such a way that they can be accessed and interpreted on current rendition systems at any time. Obviously, preserving only their bit-streams is insufficient.

improving data through migration

Migration may even improve the quality of data. Very likely, in the time span between two consecutive migration steps technological capabilities will improve considerably. Therefore, in addition to migrating a digital document into a current format, transformations that enhance the document's content may be applied. For instance, background noise can be removed from scanned images or texts. Another possible improvement would be automatically extracted indexes for text documents. In the same vein, we might add metadata for better access. Adding metadata is a topic in Sect. 3.3, where library and archive aspects are considered.

3.2 Migration in Long-Term Preservation

differentiating migration strategies

Simply speaking, "migration in the strict sense" transports an digital document from some old platform onto a new platform. There is a natural differentiation between hardware and software transfers, as well as between approaches that result in changes of the data-holding devices, the data formats, or the logical data structure. At first, we concentrate on the potential *target data formats* of a migration, especially focussing on long-lasting standards. Independent of the strategy used for long-term

preservation, metadata have to be kept permanently readable. This is also discussed in Sect. 3.3.

3.2.1 Target Data Formats

proprietary and open formats

Software variety results in a never-ending proliferation of different data formats. Many data formats are vendor-specific and depend on proprietary programs and computer environments. Prominent examples of this kind are the Microsoft-Office data formats and the WAV-Audio data format. Another category are data formats whose internal representation format has been disclosed to the public. For such *open formats*, many applications provide import and export filters. Therefore, they are important intermediaries for data exchange among heterogeneous systems. RTF and PDF belong to this category.

standardized formats ...

The growing globalization and the increasing Internet connectivity increase the demand for common data formats and communications standards. In contrast to the 1970s and 1980s, we now see enormous effort being spent on related standardization activities. Their own well-understood business interests force companies to collaborate in a growing number of standardization committees. A good example is the field of Geographic Information Systems (GIS), where there is an attempt to minimize data conversions and administration costs through the use of widely accepted data format standards. Another example is e-Commerce, where efforts concentrate on homogenizing different existing approaches to secure data transmission and jointly developing standardized data exchange formats.

... and their advantages

By introducing standard data formats not only dependency on proprietary hardware and rendition systems is reduced; also, the number of data formats that have to be administrated is reduced drastically. In this respect, standard data formats outperform all other hardware- or software-dependent data formats simply because their number is smaller. It makes sense that libraries and archives select only a few standard data formats (for each kind of document) as their target data formats for migration. This reduction in the number of target data formats helps these organizations to concentrate their know-how and thereby considerably reduce personnel costs. At the same time, migration intervals tend to become longer since standard data formats will probably be supported for longer periods of time.

formal criteria

In order to enjoy all these benefits, we propose to select data formats according to the following formal criteria:

- The data format should be public, i.e., with a full disclosure of its syntax and semantics.
- The data format should be standardized by a reputed organization such as the *International Organization for Standardization* (ISO),

American National Standard for Information Sciences (ANSI), or the *World Wide Web Consortium* (W3C). Although PDF is not standardized in this way, it has become a de facto standard for the exchange of printable documents. While standardization supported by an organization is not a must, it brings about a broad consensus among developers of rendition systems and reduces difficulties evoked by vendor-specific subtleties.

– The data format should be generally accepted and widespread. This guarantees availability of suitable rendition software on mainstream hardware platforms now and in the foreseeable future.
– In addition, the data format should be available free of patent and license fees. Organizations should be careful when selecting a data format for which the owner has reserved his rights to eventually put fees on its use. Once a format is established this can have drastic consequences as we have seen with JPEG some 15 years ago.

selecting formats

If the formal criteria are met by competing formats, the choice should be based on the organization's overall objectives and the contents of its archive. Here, options range from a simplistic perspective, where the most basic format is used for archiving (not making great demands on the retrieval system; see also Hedstrom 2000) to a more holistic view that attempts to provide the content as true to original as possible over a very long period of time. The latter option is relevant for libraries, archives, and museums which intend to conserve nondigital objects using digitalization.

document categories and data formats

Hendley (1998) has analyzed a large number of document categories and their corresponding data formats. Table 3.1 lists the major results naming current standards. For some documents containing data of a single type only, a few well-established standards exist. Two particularly successful formats for the visual representation of texts and graphics are PDF and PostScript. For compound documents (i.e., documents containing data of different types), standard data formats are still missing.[2] In this field, which among others includes virtual reality systems, Office applications, and geographic information systems, proprietary formats flourish. Some of these systems offer export filters to open up ways for data exchange (Jones and Beagrie 2001). As long as there are no standard formats for a compound document type, a document's components have to be exported separately.

problems with hypertext links

Hypertext documents, and in particular their popular linking mechanism, raise another issue, i.e., how to uniquely identify another documents (or a place within another document). For identifying a link

[2] The new OASIS Open Document Format (ODF) might become such a standard; see http://www.oasis-open.org/committees/download.php/12572/OpenDocument-v1.0-os.pdf.

category	data types	standard formats
data	alphanumeric data	PDF, PostScript, ASCII, CSV, SQL
structured text	alphanumeric, image references, markup	PostScript, PDF, TeX, DSSSL, SGML, HTML, XML
office documents	alphanumeric, bitmap and vector graphics, animated graphics	PostScript, PDF, DSSSL, RTF, ASCII, SGML, TIFF, CGM
design documents	bitmap and vector graphics, alphanumeric data	HPGL, PostScript, EPS, DXF/DWG, IGES, CGM, TIFF, ASCII/RTC
presentation graphics	bitmap and vector graphics, alphanumeric data, animated graphics	PostScript, PDF
image	bitmap graphics	PostScript, PDF, TIFF, GIF, JPEG
audio	audio data	MPEG-1 audio layers 1/2/3, MP3, MIDI
video	video data	MPEG-1, MPEG-2, MPEG-4
geographical data (GIS)	bitmap and vector graphics, alphanumeric data	PostScript, EPS, HPGL, TIFF, ASCII, CGM
interactive multimedia	all	MPEG-1, MPEG-2

Table 3.1. *Document categories and standard formats according to Hendley (1998)*

target, in today's Internet a so-called URL is used, which essentially describes how to find the target's location within hyperspace. Due to the permanent reorganization of this hyperspace, the location of a document is volatile, i.e., subject to change without notice. What is needed, however, is a more stable and more persistent unique identifier. Research is active on implementing techniques for world-wide productive use of either a *Persistent Uniform Resource Locator* (PURL), a *Digital Object Identifier* (DOI), or a *Uniform Resource Name* (URN). In Sect. 5.4.2 we will discuss these alternatives in some detail.

3.2.2 Digital Media Migration

case study: migrating sociological research data

Different motives call for migration of digital media: On one hand, the durability of digital media is inferior to that of conventional, nondigital media like paper and microfilm, whose shelf life is measured in hundreds of years. On the other hand, technological innovation in rapid succession produces even better storage devices, accompanied by new digital media of drastically improved capacity. Consumers absorb increased bandwidths, increased memory capacities, and increased access speeds like dry sponges. Typically, devices and media are replaced by better ones within a few years. They then become obsolete, since for lack of financial incentives their production and maintenance are discontinued. Soon, disproportionate effort is required to connect the unsupported devices to up-to-date computers and thus access the data stored on old

media. A case study by Green et al. (1999) illustrates the difficulty of making data from such a obsolete system available. The purpose of this case study was to migrate punch cards containing sociological research data into a hardware and software independent format.

simple and cheap

For devices and media that have not yet run out, media migration is simple and cheap (TFADI 1996). If complex structures are to be migrated or if security and authenticity issues are raised, things become more difficult and more expensive. The following discussion of concrete strategies for media migration is based on the classification from CCSDS (1999).

Refreshing Digital Media

refreshing the original or creating a replica

The simplest strategy is to refresh digital media, i.e., to create a perfect replica without changing either the format or a single bit of the content. In many cases, refreshing in the true sense (that is, refreshing the original) is not possible. Write-only CD-ROMs and DVDs are examples of that. Since a computer cannot tell the difference between the original and a replica, this is no problem. Refreshing is a preventive action taken in order to avoid losses of information due to physical effects, like de-magnetization of tapes. If properly organized and used in conjunction with redundant copies, refreshing will rule out information losses almost completely.

The TFADI definition of migration shown on p. 32 goes on to distinguish refreshing from migration in a broader sense:

> ... Migration includes refreshing as a means of digital preservation but differs from it in the sense that it is not always possible to make an exact digital copy or replica of a database or other information object as hardware and software change and still maintain the compatibility of the object with the new generation of technology.

Migration to Other Digital Media

problems although logical structure is preserved

If data are copied onto media of a different kind (for instance, from a tape to a DVD), the internal structure of the new media will often differ considerably from that of the original. Whereas a magnetic tape is *physically* organized into sequence of blocks providing sequential access to bytes streams, an optical disc is organized into sectors and blocks that can be accessed directly. Generally, different physical organization structures are handled by driver software, i.e., by part of the system software. One of the tasks of a computer's operating system is to hide such physical differences from users and application programs by providing a uniform *logical* interface for file access. The migration strategy, which

copies data between media that are of different physical, but of the same logical structure, is called *replication*. An example would be migration between a hard disk and a USB stick.

authenticity features and copy protection

When using replication, restriction mechanisms will carry over to the new media without problems if the information used by the mechanism (e.g., checksums) does not depend on physical properties of the accessed media. Other authenticity features (e.g., digital watermarks, digital signatures) or strict copy protection directly depend on properties of the data representation or of the physical media themselves. In that case, media migration is severely restricted. Most movie DVDs and a growing number of audio CDs are prominent examples. Here, media migration requires that the archiving organizations (libraries, archives, and museums) and the producer of the media cooperate – if only for legal reasons.

remove copy protection and digital signatures prior to archiving

This kind of cooperation is also needed if copy-protected or digitally signed documents are submitted to a public record office. Here, we recommend that original copy protection and digital signatures be removed (and, possibly, be replaced by corresponding mechanisms under the office's control) when preparing the documents for long-term archiving. Of course, this action has to be documented accurately and in full detail in order to make its effects traceable to future users.

increased storage capacity

So far, for archiving purposes mostly slow but high-capacity media like magnetic tapes and optical disks have been used. Prices for hard disks have fallen to a level where they can be considered a fast and reliable alternative to classical media. This would bring a new quality to archiving since hard disks – in contrast to magnetic tapes, etc. – support direct on-line access. The advantage is even bigger if access can be provided world wide over the Internet. Among other new technologies, DVD systems are about to displace CD technology in the IT market. In combination with a robot exchange system, DVD systems offer enormous storage capacities, and, if connected through fast bus systems (e.g., SCSI), allow for data access as fast as their hard disk counterparts.

3.2.3 Migration to Nondigital Media

a grim scenario for retro digitization

According to Hedstrom (2000), copying digital information to nondigital media is one of the most common and widely used strategies for long-term preservation. This is quite remarkable in the presence of huge retro-digitization projects conducted by libraries and museums all over the world in order to preserve historical documents and other artifacts. If we do not succeed in establishing an effective long-term preservation strategy, these digital data may well become useless (or inaccessible) long before the destruction of their historical originals.

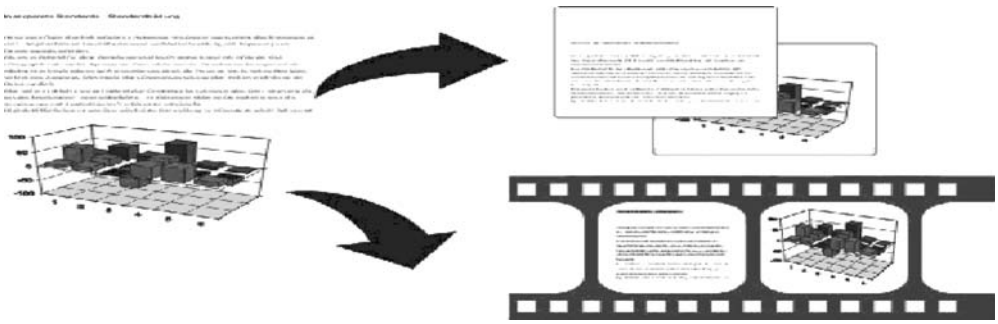

Fig. 3.1. Migration to nondigital media

paper and microfilm

As illustrated in Fig. 3.1, this archiving strategy makes use of two media, namely paper that carries the printed digital data, and microfilm that stores photographic copies of the original. The old idea to store digital information (for instance, as bar codes, or as glyphs, or in any other encoding) directly on the microfilm is strictly limited by the resolution of the film. Due to this limitation, at best some metadata can be stored digitally. The advantages of paper and microfilm are obvious: They resist to many environmental influences and deteriorate only slowly over time. Also, no expensive special-purpose hardware and software is needed for transferring digital data to paper or microfilm: a commercially available printer will do. The invaluable pro is, however, that their reproduction is simple and comes without expensive and failure-prone interpretation system – the human eye and a magnifying glass is all that must be at hand. Simple text and unstructured documents are ideal candidates for this kind of archival strategy.

no full substitute for digital media

Unfortunately, paper and microfilm are not suitable for complex and compound documents such as databases, videos, or hypertexts. There is no direct way to transfer these data to nondigital media adequately. The same is true for spreadsheets and other computationally enhanced documents. As soon as a table, for instance, is printed to paper, the formulae behind the numbers in the table are lost. Documents are "flattened" (TFADI 1996). Safety and security protocols, all authentication mechanisms, digital watermarks, etc. are lost as well (unless special care is taken to add them as binary coded information).

Microfilms

an established and mature technology

Microfilms are a wide-spread class of nondigital media that are fit for archiving purposes. Although micrographics is not suitable for all types of digital objects, it offers some attractive features. As already mentioned, using simple physical devices their contents can be made directly perceptible to the human eye. For the analog image

representations stored on them, no decoding of data and metadata that makes the digital world so tricky is needed. Micrographics is an established and very mature technique. A high confidence in microfilms results from long-standing experience of professional archiving institutions and service bureaus. Formal standards,[3] recommendations, and guidelines facilitate archiving. Compatibility is no severe problem. Information exchange among different systems is easy. Most importantly, the imposing durability of these storage media significantly reduces the frequency of migration steps. In the digital world, we see a need for migration in very short-time intervals – hardware and software are evolving fast. In contrast, microfilms promise a lifetime ranging from at least 50 years (if standard DIAZO films for every-day use are employed) up to 500 years and more (for high-tech silver-gelatin-type microfilms). Even if these ratings would be considered too optimistic by some, extensive tests, defined quality procedures, and the proven readability of films that are over 100-years old have already produced a high degree of confidence.

microfilms offer unrivaled advantages

Microfilm is a suitable medium for storing particularly sensitive data. For some considerable time now, long-term archiving of *analog* images as described above has been established as a reliable technique. However, information can also be placed on microfilm as purely *digital* data! This way, significantly higher storage capacities can be achieved than with analog imaging. Analog photography of the Bible would require 30 m of film; as digital data the same amount of information would require only 5 mm. At the same time, the digital data are preserved with the highest standard of security. Microfilm offers unrivaled advantages as a storage medium in terms of durability and protection against forgery. For example, it is possible to scan in documents bearing important signatures and record them as analog images – possibly supplemented by digital metadata to facilitate retrieval. With microfilm, long-term storage relies on conversion. This means that the relevant data are transformed into a format that can no longer be altered and is easy to access. The data are thus preserved on the medium once and for all – viruses stand no chance. Since the production of films is relatively cheap, additional security can be achieved through redundancy by storing backup copies in distributed locations.

archival storage: working version and microfilm in parallel

Thus with regard to security, microfilm is the high-end version in a graded concept of storage media. And in terms of cost, too, it can certainly hold its own with other media. At any rate, relevant costs are only incurred in its manufacture and in data retrieval. For storing the films, archivists have virtually no expenditure (measured in terms of storage

[3] Issued, in particular, by the American National Standards Institute (ANSI), the Association for Information and Image Management (AIIM), and the International Organization for Standardization (ISO).

capacity). However, for daily, rapid availability, microfilm is not suitable. In parallel to the back-up copy on film, quickly accessible working copies should be provided. In combination with microfilm, there is no need for the on-line versions to meet any special security standard. If the data on the working version should damaged or even destroyed, the microfilm copies can be retrieved. For data that are only needed in rare, exceptional cases, on-line versions are dispensable. Examples are personal data, medical reports, and documents relating to liability or court files.

micrographics in detail: recording and reproduction

In order to convey an idea of micrographics' potential we describe some technical features and projects. Traditional micrographics are fully processed in an analog way. Analog cameras mostly take the pictures of paper-based material. *Planetary cameras* are the right choice for brittle and fragile documents as well as for bound volumes, which are placed on a flat copy board. Typically, these cameras support 11 and 35 mm films. This recording technique produces high quality images. If lower quality is acceptable and high processing speed is required, *rotary cameras* are the adequate equipment. The rotary motion of the copies, which is synchronized with the film's rotation, requires flexible and uniform-sized material. This technique is generally limited to 16 mm films. Variants of planetary cameras, so-called *step-and-repeat cameras*, produce *microfiches*. They consist of a grid of pictures, which is usually placed on a section of 105 mm microfilms. Microfilm and microfiche readers are optical devices used for viewing and/or printing. *Computer Assisted Retrieval Systems* (CAR) allow for the retrieval of indexed documents. They usually require 16 mm films housed in cartridges. Opaque marks on the film, so-called *blips*, and frame numbers help to locate sought-after exposures.

on choosing the reduction ratio

In order to be able to reconstruct a quality image of an original picture, a suitable reduction ratio must be chosen for recording. Microfilmed information very often consists of discrete symbolic formats like texts, architectural plans, or engineering drawings. The legibility of such information provides a clear quality measure. Often, the following reduction ratios are used: Textual business documents in legal or letter size format fit on 16 mm films. If copies exceed the 11–14 in. format, a 35 mm film is recommended. This film width avoids reduction ratios above 32. Libraries prefer ratios in the range from 10 to 20, whereas the ratios of technical drawings, which are often recorded on microfiches, range from 24 to 40. In addition, convenient handling, space-saving use of films, and efficient generation of copies determine the optimal reduction ratio. We must also keep in mind that even small impurities, e.g., so-called microspots, can possibly destroy important information.

storage of films

Note, however, that such long periods of time require a careful storage of films. Of course, the development and handling of films according to the manufacturer's specifications are vital preconditions.

Humidity, temperature, and their changes, as well as the cleanness of the environment influence the mechanical and chemical stability of a film significantly. Using the central archiving of microfilms from the federal and regional archives of Germany as an example, we show how to guarantee optimal and disaster-safe conditions over very long time. Kegs (slightly modified versions of kegs used by beverage industries) serve as containers for film spools. The kegs are stored centrally in an ancient silver mine, which looks like a cellar of a brewery (Fig. 3.2). This location provides an almost constant temperature without permanently consuming energy. The first major step is to adapt both the film spools and the kegs to silver mine's temperature. For this purpose, a climate chamber is used (Fig. 3.3). After the films have become acclimatized, which takes about four weeks, they are canned. Currently, there are two tunnels housing some 1,700 kegs containing silver films with a total of about 1,000,000,000 pictures. While originally the Cold War fostered this bombproof solution, since then rampant fires in historic buildings (such as the 2004 conflagration of the Anna Amalia Bibliothek in Weimar, Germany) have demonstrated the necessity of having a safe – and cost-efficient – location for extensive backups.

Binary Coded Information on Microfilms

Computer Output Microfilm

The above discussion refers to the analog sides of micrographics. But what about digital material? *Computer Output Microfilm* (COM), a technique already wide-spread in the early 1980s, skips printing to paper. Film recorders directly transfer computer files onto microfilms and microfiches – also referred to as *digital films*. The first applications primarily aimed at an inexpensive distribution of large quantities of

Fig. 3.2. *Storage of kegs*

Fig. 3.3. *Climate chamber for microfilms*

data-oriented (alphanumeric) documents such as parts lists. Nowadays, the enhanced quality of recorders (and scanners, if no digital sources are available), combined with the increased processing capabilities of computers, allows raster images of high resolution to record at an impressive speed. Many archives and libraries, including, e.g., the Cornell University Library (CUL),[4] use this method for archiving purposes. The digital processing also allows different quality on-line versions to be generated. These may serve, e.g., as low-resolution working copies or finding aids (thumbnails).

color laser: devices and films

Technologies developed for the motion picture industry further improve the production of digital films. The ARRILASER is an example for this kind of synergy.[5] An adapted version of ARRILASER is used to transfer high-quality images to 35 mm color microfilms for preservation purposes. Three solid-state lasers, one for each basic color, i.e., red, green, blue (RGB), replace the traditional Cathode-Ray Tubes (CRT). Figure 3.4 shows the laser optic unit. The powerful laser light makes it possible to use fine-grained film material, which is inherently less sensitive to light (hence, more durable). The stability of the light ray and the small diameter of the ray further improve the quality of exposures. However, watching aged color films may cause some doubts about the fidelity of colors. Indeed, there is no broad market for durable color films. Fortunately, Ilford offers a high quality color film, named ILFOCHROME MICROGRAPHICS

[4] Digital to Microfilm Conversion: A Demonstration Project 1994 – 1996, see http://www.library.cornell.edu/preservation/com/comfin.html.

[5] Fraunhofer IPM in cooperation with ARRI, see http://www.ipm.fraunhofer.de and http://www.arri.com.

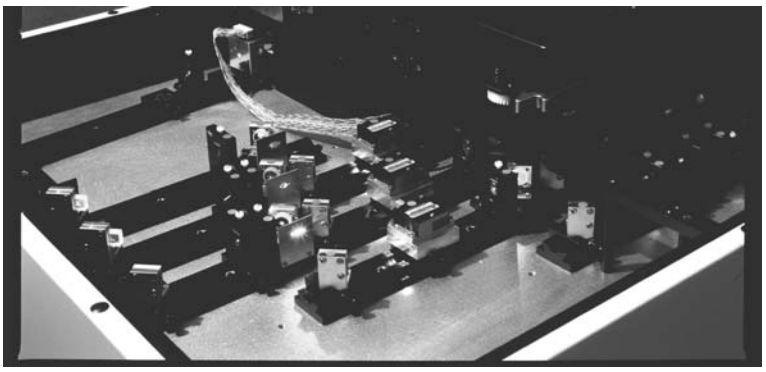

Fig. 3.4. *Laser optic of the ARRI film recorder*

(previously CIBACHROME). Here, a silver-dye bleach technology provides the basis for long-term fidelity of colors: During manufacturing, stable colors are embedded in the emulsion layers of the film; the *development* process, in particular the bleaching, brings out their visibility.[6] Due to this special development process and other measures the durability of ILFOCHROME films rivals that of black-and-white films – which is more than 500 years.

archival storage: storage capacity

Fraunhofer IPM is working on a storage concept providing an extremely high data storage density that is achieved through the digital exposure of films. The exposed dots, as produced by the adapted ARRI-LASER, measure about 3 µm. Accordingly, about four gigabytes of data can be accommodated on 1 m of 35 mm film. This is equivalent to the capacity of five CD-ROMs. A single reel of film would be sufficient to store the entire data stocks a medium-sized enterprise produces within one year.

Archival Storage: durability of media

After processing (as described above), the microfilm ILFOCHROME MICROGRAPHICS fully complies with the requirements for internal auditing standards. Even under extreme climatic conditions, color microfilms display their stability. In tests at 80°C and with high humidity no signs of degradation in the color were observed whatsoever. These tests suggest that the film has a shelf life of more than 500 years. A time scale such as this is particularly relevant for preserving cultural assets, which has been one of the prime uses of this material so far. As indicated, adequate conditions for physical storage of the films must be ensured. Both in Europe and in the USA, service providers maintain special underground caverns for long-term archiving.

Accessing Information on Microfilms

access to analog documents on microfilms . . .

If microfilms are considered an interesting alternative to digital forms of long-term preservation, how can we guarantee access to complex

[6] For detailed information, see the specification at http://www.ilford.com.

document collections without continuous recourse to computer-based systems? A solution developed by the General Directorate of the State Archives of Bavaria for governmental and administrative records (Luprian et al. 2004)[7] aims at both, low costs for long-term storage, and re-use in future computer systems.

using directories and metadata ...

A general supposition (that can be taken for granted in practice) is that collections of archiving units – digital files and, likewise, physical units – are organized in hierarchical tree structures. Computer file systems can manage such structures in a natural way: In the directories of computer file systems, physical units are represented by digital files containing *archival reference numbers* for retrieval and other metadata. The capability to include links to other tree structures of the same kind (comparable to the traditional *mount* feature in the Unix operating system, where entire file systems can be glued together and accessed homogeneously), as well as the capability to assign arbitrary metadata to files are features worth to be sustained. On the other hand, the complexity of current computer file systems may hamper long-term preservation, and also their built-in capabilities for expressing metadata are deemed insufficient.

stored on microfilms ...

Therefore, the next step is to represent these structures independent of the concrete file system structure of any particular operating system. Using a simple markup language (a restricted form of SGML) the hierarchical directory structures are represented in a computer and human readable text format. All relevant metadata are contained in named metadata text files. Such a set of files can be stored in a very primitive file system, even more primitive than *tar* archives as known from magnetic tapes. Reconstruction of the original structures is straightforward, if files are retrievable by their names. The complete primitive file system is stored permanently on microfilms, where each (directory or metadata) file is represented by a named sequence of one or more images containing text.

and rebuilt using OCR

Optical character recognition (OCR) can rebuild these files reliably if unambiguously interpretable fonts are used. Thus, the aims of low-cost long-term storage and readability by future computer systems are achieved. Beyond that, scanning can even rebuild graphical image files to some degree. Bitwise identical copies w.r.t. the original digital representation cannot be guaranteed, of course. However, considering the many advantages of microfilms, this restriction may be acceptable for a lot of originally "unstructured" document types.

access to binary coded documents,

Similarly, in the Fraunhofer IPM system the reading process is designed as an open system, meaning that instructions on how the encoded (binary) data are to be interpreted are saved on the film as meta-

[7] A preprint is available on-line: The key role of metadata for permanent preservation of digital records in the archival environment – A message from the far future, see http://www.gda.bayern.de/AGTEXTE/marburg.pdf.

preservation metadata

data. To ensure that the data can still be read when the film has seen several generations of equipment and software come and go, the software and equipment configurations are written to the film as well. An eye-readable, analog representation of metadata offers a secure entry point to the *representation network* (see Representation Information as defined in OAIS). This bootstrapping approach results in autonomous, complete, and physically coherent storage capsules (AIP). As any other physical equipment, microfilms and its processing devices are not perfect. This requires additional precautions, e.g., redundancies, in order to be able restore the correct bit-stream reliably.

3.2.4 Migration by Transformation

transformation

Various other migration strategies apply transformations that affect the logical structure of a document. These strategies are subsumed under the notion of *content migration*. One content migration strategy transforms documents into standard formats supported by current application programs. This strategy is important if documents have to be kept both readable and manipulable all the time. Another content migration strategy will compress the data in order to save storage (using, for instance, a zip program). For restoring the original format, an inverse decompression program (say, unzip) must be available. These tools are said to be *lossless*, if the transformation sequence zip–unzip, when applied to any document D, will restore D in its original format. Note, however, that restorability of compressed data critically depends on the availability of suitable de-compressing software. This makes long-term preservation of a rendition system with an integrated compression facility rather complex and cumbersome: The de-compressing software must be part of future migration steps, or, if storage space should eventually be no longer a problem, all data have to be uncompressed in time.

upgrading

The rapid changes in software together with the increasing number of new standard data formats are the reasons for ongoing data transformation. Most software products provide easy and simple ways to "upgrade" documents, i.e., to convert documents stored in the data format of an old software version into the current format.

migration losses

A transformation that converts data from one format to another always tries to preserve logical structures, access restrictions, formatting, and other relevant properties as true to original as possible. Still, losses are likely and in many cases unavoidable. Even a migration of ASCII documents might prove problematic: Whereas in a simple text document, the number of syllables per line is irrelevant, in lyrics the segmentation into verses and lines, and also their formatting are aspects intrinsically tied to a particular poem. Also remember the line graphics example of Sect. 1.2, where readability of the document crucially depends on the conservation of certain typesetting features (carriage

returns, fixed width font, and size). In the following we discuss some tool-supported migration techniques.

Upgrades within a Product Family

compatibility within product families

In software industry, nowadays almost every new release of a software system comes with a plethora of new features. As a consequence, their data formats often need to be extended. Prominent examples are Microsoft Office products. Naturally, users want to be able to edit old documents with new versions of the same software product. Typically, this is handled in the following way: If an old format document is opened, the software offers to convert the document to the new format, which then can be saved. That way, no information is lost if the new format is "downward compatible" to the old one. Since software developers can not at all guarantee this downward compatibility for an extended period of time – nor even the mere future existence of their enterprise and its products – this kind of migration technique alone is not sufficient. Other, mostly financial disadvantages result from the fact that for all data formats at least one application must be preserved and regularly migrated. Microsoft Word, for instance, has seen more than six format changes within ten years going from Word 6 to WinWord 2000.

Exports into Foreign Formats

indifferent support of competing formats

Another frequent option in today's application software are *export filters* allowing documents to be transformed into the data format of another product. The fact that such filters are provided for wide-spread target formats only, is of no concern for long-term preservation, where insignificant proprietary formats play no role. For obvious reasons, customers cannot expect that software developers invest a lot into export facilities for foreign and sometimes competing formats. Exceptions are product-independent (de facto) standards such as RTF and PDF, or markup languages such as SGML or XML. Some newcomers try to gain increased market penetration by pretending compatibility of their products. Often, such companies do not fully disclose the specification of their formats. If using export filters as migration tools, information losses must be tolerated up to a certain degree since almost invariably the potentials of original and export formats will differ, if only in minor aspects. Whether these differences are deemed important, depends on the end users and on the tasks and the particular archiving strategy of the library, the archive, or the museum.

Import of Third-Party Formats

direct and indirect imports

For the import of documents, i.e., for transforming foreign formats into an application's native data format, often so-called "import filter" tools

are provided. From a software vendor's point of view, import filters are more attractive than export filters because they support user transition *towards* their own product. Like export filters, import filters cannot guarantee lossless migration. If two application software systems are not directly *interoperable* at all (i.e., there are neither direct import nor direct export filters applicable in either system), then as a last resort one can migrate a document via a third intermediate format. Since this involves two distinct migration steps (from foreign to intermediate and from intermediate to native format), the risk of losing information is doubled.

Application of Conversion Software

commercial conversion tools

In the market, a variety of commercial tools for data format conversions are available for sale. These tools collect a bunch of import and export filters for the most attractive source and target formats in a single product offer. For m source formats and n target formats we have a total of $m \times n$ possible conversions, that each have to be programmed individually. In order to reduce the effort involved, most products seek a remedy in the use of a stable intermediate format. Such an intermediate format reduces the programming effort to only $m+n$ conversions to be implemented – m conversions from the source format to the intermediate format, and n conversions from the intermediate format to the target format. This, of course, again doubles the risk of losing information.

3.3 Archiving Processes of the Migration Approach

In the following, we describe how the migration approach can be embedded into the organizational framework introduced in Chap. 2. In particular, we map migration activities to archiving processes defined in the DSEP model and the underlying OAIS Reference Model, respectively.

the goal

The overall goal of all archiving activities using the migration approach is to ensure the following: At any given time, for each archived document there exists at least one up-to-date hardware/software platform. These platforms may (and will) differ from the platform running the archival system.

Should we keep old versions?

Falsification is a general risk of transforming migration. If huge amounts of documents are automatically transformed, even sophisticated quality control may overlook migration errors. Should we, therefore, keep old document versions for safety reasons – even though it will be impossible to render these document versions in the future? Consequently, old versions should be preserved only for important documents, for which a future reconstruction seems at least plausible. Let us, however, not underestimate the intellectual capabilities of future "data archaeologists" who will attempt such reconstructions.

needed software

Typically, the rendition system for a particular document comprises two software components (a) the operating system of the current hardware platform and (b) the application program used for viewing, processing, and editing the document. Note that the operating system (e.g., Windows XP) is the most important prerequisite for running the application program. In contrast, the type of the underlying hardware platform (e.g., PC by IBM or Siemens or any other vendor) is of secondary importance only.

platform and environment

Description: Below, we will use the terms "computer platform" and "environment" as follows: A compatible combination of computer hardware and operating system is called a *(computer) platform*; a compatible combination of platform and application program is called an *environment* (for a class of documents).

The definition of a set of *standard environments* for all supported document classes is one of the most important decisions for long-term preservation management. Moreover, this decision must be continuously revised and adapted to the current preservation purposes.

downward compatible product families

Over time increasingly powerful and better performing versions of computer hardware, operating systems, and application programs will be developed. We call a series of product versions where each member covers the complete feature set of all previous releases (often, more) a *downward compatible product family*. Given a downward compatible product family, for long-term preservation purposes old family members can be replaced by new ones. Older members can still be kept for some "important" product families if future reuse seems plausible.

the role of metadata

Metadata describe different aspects of a digital document. This includes technical features and information about selection criteria such as bibliographic information, key words, archival numbers, and many more. Metadata play a double role: On the one hand, metadata are important for describing the data of a document and, therefore, belong to its data capsule. Consequently, the archival system regards metadata as an integral part of the document *content* and stores them accordingly. On the other hand, the archival system needs metadata for document accesses – they are relevant *catalog data*, too. For reasons of efficiency, metadata are often stored (redundantly!) separate from the document proper in a format, which is optimized for fast access, e.g., in a database. This double storage of metadata in possibly two different formats raises consistency issues. The problem of guaranteeing consistency among different metadata representations becomes a true challenge if over a long period of time a series of migration steps has been performed. Whether to store metadata in two formats or in one format (and deducing the other format on request) is a very delicate decision best left to vendors of archival systems. Since long-term preservation

is the overall goal (and "long-term" means *long* term), robustness must take priority over efficiency.

metadata formats

In most cases, metadata consist of structured text. We, therefore, recommend structured text formats such as SGML or XML; see Sect. 6.1 for details. Since structured text formats (SGML or XML) are usually stored using basic text formats (ASCII or Unicode), migration of metadata should not be very challenging even in the long run. For the preservation of metadata, the same conditions hold as for the rest of the data capsule.

Referring to the DSEP process model as illustrated in Chap. 2, the following tasks have to be considered in a migration context: *capturing* a new digital document, *preserving* an digital document, and *accessing* an digital document.

Capturing a New Electronic Document with the DSEP Processes *Delivery & Capture* and *Ingest*

Is the rendition environment available?

Typically, a new digital document to be ingested belongs to a class of documents, which can be rendered using one of the standard rendition environments that are already supported by the library (according to an earlier decision of the library's management). In that case, it suffices:

- To check the document class (this can be a selection process, if the document matches several supported classes) and its suitable environment (this can also be a selection process, if multiple suitable environments exist for the selected document class).
- To capture the corresponding metadata, which provide the usual bibliographic entries and links to all known environments that may be used for rendering the document (where environments supported by the library should be distinguished from other environments).

environment not yet available

If a new digital document cannot be handled by any one of the defined standard rendition environments, then the capture process for this new document is adjourned. The issue is referred to library management, which decides about further processing.

- If library management considers the document to be a first and interesting candidate for a new document class, a new standard rendition environment has to be chosen and added to the set of already maintained environments. After that, capture of the new digital document continues as described above. Note that the decision to support another environment may be costly.
- Otherwise, the capture process for the new document is canceled completely.

Preserving Document Repositories with the DSEP Processes *Archival Storage* and *Preservation*

automatic refreshment

Like for all other archiving strategies, the data capsules of all digital documents need to be refreshed periodically in order to avoid demagnetization and other wear outs. This is why the OAIS Reference Model defines refreshment and replication as important migration steps (cf. Sect. 3.2). An archiving system must perform these steps regularly and automatically. In this context, additional safety measures like making backups and replicas must be integrated.

packages remain unchanged

If the composition, the contents, or the archiving location of an Archival Information Package (AIP) is about to change, another kind of migration step called repackaging is required. However, neither OAIS nor DSEP allow for changes in stored AIPs. Instead, such an AIP has to be extracted from the archive as a normal DIP (Dissemination Information Package), which can then be modified as needed. Afterwards, the modified package is re-submitted as SIP (Submission Information Package), just like any other new package. We hope that for trivial changes the archiving systems will provide more pragmatic and simpler solutions. The following measures are very expensive because they potentially have impact on every single document within the archive.

permanent monitoring of new technologies

In the DSEP model, a subprocess of the *Preservation* process includes an activity called *Monitor Technology*. This activity permanently watches technological developments (hardware, software, storage media) and proposes suitable reactions to library management. For the migration approach, the following events should trigger library management decisions:

adding a new environment

– *A new environment proves relevant*. If a new environment is added to the set of standard rendition environments maintained by the archive, then, for every document, it must be checked whether it is supported by the new standard environment. If so, the metadata of this document have to be updated correspondingly.

Trust but verify!

In case a new standard environment contains new version of the operating system software or of the application software, then these new versions need to be included as new AIPs. If the vendor guarantees downward compatibility, previous versions may become dispensable. Despite these guarantees, we recommend, however, to perform at least some spot checks on the own document corpus before replacing software versions previously used.

standard environments running out

– *Part of an existing standard environment becomes obsolete*. All rendition environments containing this component have to be removed from the set of supported standard environments. After that, every document be must checked whether it was supported by a removed standard environment. If so, the metadata of the document must be

updated correspondingly. In case the removal makes a document inaccessible (because there are no more environments supporting the document), the resulting loss of data may be avoided through the following measures

– A new environment supporting the otherwise inaccessible document is found and added to the set of standard rendition environments. Typically, this happens whenever an operating system is replaced by a new downward compatible version. *find a substitute environment*

– If in the near future the rendition system's software for the document will not be supported any more by any remaining environment, then the document must be migrated into a new format that can be rendered by software running on a current standard environment platform. This conversion may prove easy, if the old document format is one of the import formats of the chosen application software or if suitable conversion tools are available. Conversion is, however, more challenging if the document formats of the old and the new environment differ to a great extent. In this case, information losses become very likely. Consequently, library management should focus on proactively avoiding this scenario. If, however, this situation cannot be avoided at all, each conversion step must be carefully documented within the metadata of the migration history. In addition, the old document version should be preserved. This hopefully enables later generations of scientists to draw conclusions and inferences about the original document content, and possibly even restore it after a period of inaccessibility. *perform a migration*

Accessing a Document with the DSEP Process *Access*

Accessing an digital document using the migration approach is simple and easy. Recall that at any given time, due to migration, the original documents are kept in formats that are supported by a current standard environment. *simple access*

If, however, authenticity of a document is of major importance, and multiple transforming and structure-changing migration steps have been performed, the migration approach may become problematic. Here, the only remedy can come from (a) sufficient metadata within the migration history, (b) storage of old versions of the document, and (c) a sophisticated inference scheme for reconstructing the original document content. Unfortunately, this reconstruction will not only be very costly, but also may fail completely. *But is it authentic?*

3.4 Strengths and Weaknesses of the Migration Approach

strengths

In the earlier sections, we already have discussed the strengths of migration in detail. Here, a quick summary:

documents are always accessible

– The migration approach enables us to access up-to-date documents at any given time. Required hardware and software are always available within standard environments. In contrast to the emulation approach, end users can use *current* rendition systems. Consequently, they do not or need to learn old, cumbersome, and sometimes unergonomic features of "ancient" interfaces. Given the enormous innovation speed in the IT industry, it is hard to imagine that ordinary users would still be able to handle today's document management environments in, let us say, 100 years from now.

and processable

– But access is not the only reason for the attractiveness of the migration approach. Since all documents are accessible in an up-to-date format, functions of current tools can be applied (e.g., search facilities or cut-and-paste). Both web publishing and collaborative work in a distributed environment are easy. Today, such additional functionality is "en vogue." We sometimes take it for granted without thinking about it any more. In the future, surely new functionality will be developed. Everyone will use it – and apply it to migrated document as a matter of course.

improved presentation

– Often, migration to a current document format implies an *improvement* over the previous format. Since the days of the typewriter, a rich set of new presentation features has been developed. These features not only concern pure text layout but also sophisticated embeddings like graphics, animation, or "active text elements" (e.g., spreadsheet cells, hyperlinks, embedded JavaScript, or Active-X controls).

less training

– Accessing current document formats within up-to-date rendition systems significantly reduces training for editing the documents. The same holds for the migration steps themselves. If applied regularly, migration becomes common knowledge in the library community.

tools available

– Due to an urgent and increasing demand for conversion tools, such tools will be available for (at least) the most important document formats, and for a reasonably long span of time.

"alive" documents

– Continuous migration into ever new document formats requires us to permanently revise the whole document corpus. Otherwise, the quality and possibly even the meaning of the originals are at risk. By this process, documents are kept fresh in our minds all the time and, therefore, "alive."

weaknesses

Certainly, quality assurance (if done properly!) is an advantage of migration. Never-ending revision and *alteration* of the document corpus indicate, however, serious weaknesses of the migration approach:

ad hoc solutions

– Migration is not a single, precisely defined method. Instead, migration is a rather holistic term for a variety of very different transformation techniques. Since new document formats keep sprouting everywhere, we cannot reasonably envision one single precise and uniform method ever to work for all of them. Probably, new conversion tools will have to be developed all the time. Reliance on ad hoc solutions, however, is a serious detriment to the overall migration approach! Fair, you can use state-of-the-art tools to analyze document formats. It is, however, still a tough nut to automatically decide whether documents of different formats can be regarded semantically compatible to each other. Structural and semantic compatibility between source and target formats significantly influences the complexity of a particular migration step. This particularly holds true for compound documents (i.e., documents created in a format mix), which are – as we have seen – so en vogue today.

risk of falsification

– Transforming migration invariably carries the risk of falsification, which dramatically increases if transforming migration is applied repeatedly. Errors accumulate! Consider the line-drawing example from Sect. 1.2. Was it the wish of the creator to use text graphics, or did he use characters simply because line graphics were not available at his time of writing? If better tools had been available, what would his graphics look like? Without interviewing the creator personally – which as a rule is impossible in long-term preservation scenarios – questions of this kind will remain unanswered. As a consequence, any transformation into a true graphic format is speculative. On the other hand, even today text graphics are not considered acceptable any more. Therefore, migrators are left with the responsibility to avoid accumulating errors to the best of their (incomplete) knowledge. In this scenario, it makes sense to store the original version (in our example, the text graphics version) in parallel to every new version. Also, each alteration must be documented carefully. This problem is well known also outside the digital world. One of the most prominent "analog" examples for the above discussion are transcripts of the Holy Bible. Here, it was and it will be important to have access to previous versions in old languages. Using them, we have a chance to recover from some of the consequences of accumulated misinterpretation and error-prone translation speculations.

Is the Bible authentic?

high expenses

– At each migration step, all documents of a particular document class have to be fetched, transformed, converted, and stored again. This holds for an ever-increasing number of documents in more and more

heterogeneous formats, thus causing costs of migration to "explode" eventually. Unfortunately, even if migration were to be performed automatically, quality-assuring and restorative measures still need to be applied. Only in rare cases can these measures be automated – instead, they are personnel intensive, time consuming, and expensive. Facing the enormous data volumes that are discussed in the community, we argue that measures maintaining (or restoring) high quality will only be applicable to small collections of very precious and very important documents.

http://www.springer.com/978-3-540-33639-6

Long-Term Preservation of Digital Documents
Principles and Practices
Borghoff, U.M.; Rödig, P.; Scheffczyk, J.; Schmitz, L.
2006, XV, 274 p. 67 illus., Hardcover
ISBN: 978-3-540-33639-6