nestor

nestor / Institute for Museum Research

Nothing Lasts Forever

nestor - ratgeber 1

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Materialien aus dem Institut für Museumsforschung – Sonderheft 2

The „nestor – ratgeber“ series is published by

nestor Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen in Deutschland

Network of Expertise in Long-Term Storage of Digital Resources

http://www.langzeitarchivierung.de

Co-publisher of this volume in the nestor series is:

IfM Institut für Museumsforschung der Staatlichen Museen zu Berlin, Stiftung Preußischer Kulturbesitz

Institute for Museum Research, Germany

http://www.smb.spk-berlin.de/ifm

Authors:
Dr. Stefan Rohde-Enslin, Institut für Museumsforschung (SMB-PK)
Dr. Keith R. Allen

Photography:
Dr. Stefan Rohde-Enslin, Joshua Enslin

ISSN 1860-4641 (Materialien aus dem Institut für Museumsforschung. Sonderheft)
ISSN 1860-4706 (nestor-Ratgeber)

URN http://nbn-resolving.de/urn:nbn:de: 0008-2010031529

Contents

Preface to the 2nd Edition 2
Dear Colleagues, ... 4
It's All How You Look at It 6
File Formats 12
Formats in Detail 15
1. Image Formats 15
TIF 15
JPEG 17
JPEG 2000 19
GIF 20
Additional Image Formats 21
2. Text Formats 22
DOC 22
RTF 24
TXT 25
PDF 26
PDF/A 27
3. Video Formats 29
4. Audio Formats 30
5. Database Files 31
6. Formats, by JHOVE! AONSer is here! 32
Storage Media 33
Media … 36
Diskettes 36
ZIP Disks 38
Magnetic Tape 39
Hard Drives 41
CDs 43
DVDs 47
Blu-ray Discs 48
Integrated Circuits 49
USB Sticks 50
Summary: Storage Media 51
Further Considerations and Recommendations 53
Appendix 56
Overview of our recommendations 56
XML 57
Archiving E-Mail 59
Further questions? Consult Preservation Experts 63

Preface to the 2nd Edition

The first edition of Nothing Lasts Forever was published in 2004. Five years later, the long-term preservation of digital information in museums remains unfinished business. This fact, coupled with new developments, such as the growing adoption of PDF/A and the advent of flash memory media, encouraged us to produce a second edition of this handbook. The use of media in today's cultural heritage community defies easy generalization. Whereas many remain loyal to older media and software, others have long abandoned DVDs for the latest Blu-ray disk drives. Both of these approaches come with their share of challenges.

This second edition discusses new developments in light of difficulties raised by the evolving nature of digital media. We find ourselves repeating advice contained in the first edition, encouraging museum professionals to embrace change without chasing fads. In many respects, our "bottom line" message remains the same: unwelcome as it may sound to some, ensuring access to digital media is only possible insofar as one commits to a long-term approach to digital preservation.

This second edition also contains much new information, including a series of recommendations on e-mail archiving that acknowledge the complexity of building your own digital repository. This edition, like its predecessor, offers many practical suggestions you can enact right away, but in this new edition we also seek to take into account how nuanced full-scale digital preservation has become in recent years. As in the first edition, our message is not to despair, but to act wisely. Significant efforts are underway in many countries to tackle the problems associated with media archiving.

As we discovered in preparing this second edition, the questions we're asking in cultural heritage organizations are remarkably similar on several different continents. To both reflect and engage this discussion, we decided to publish this second edition in English as well as German. Revisions and additions were completed by Stefan Rohde-Enslin and Keith R. Allen. Another reason for this second edition is that we some time ago ran out of print copies (as it turns out, demand for information stored on paper remains vigorous as well!).

It would be nice to claim that this will be the final edition of our booklet. We'd like nothing better than to witness the problems associated with continued access to digital media vanish into thin air. After all, if nothing lasts forever, than these difficulties, too, must eventually pass.

Rest assured, a talented group of professionals is working hard to ensure that the long-term conservation of digital information becomes commonplace. In the meantime, we hope this simple guide contributes to solutions you and your organization will embrace as your own.

Berlin, 2009

Dear Colleagues,

Actually it's all pretty simple. Everything changes. Old replaces new, only a few select items survive well into the future. One of the main tasks all museums face is ensuring that collections are protected against the forces of disintegration. We strive to preserve for as long as humanly possible ideas, techniques, tools, and, above all, the objects that result from our efforts. To achieve this goal, museum professionals employ ideas, techniques, and tools. These, too, change over time, giving way to new ones.

Databases replace card files, just as image and text files replace paper prints and stationery. These changes are in some respects beneficial. Much of our work can be accomplished faster and easier than in the past. Today almost anyone can combine text with pictures, and there are now multiple ways to manage effectively large numbers of digital objects. In sum, it seems as if the digitization of the tools we use to go about our daily chores has actually made life much easier, right? At the very least, it's worth taking a closer look at them.

What has changed in our digital world is no small matter. It involves more than simply substituting older formats for new. The change is actually more fundamental. More and more of what we need to go about our day-to-day activities is available only in digital form. The advantages of this, we acknowledge, are immense, but the disadvantages should not be overlooked. To accomplish our work in museums and other cultural heritage solutions, we have become dependent on computers: whereas information from index cards can be read with the naked eye, information from databases can only be recalled and interpreted with the aid of technology. As a thought experiment, just imagine what it would be like if the computer you work with were to disappear from your desk. Can you recall which bibliographic references you assigned to the object with the inventory number of 1014?

Like we said, everything changes. As you can imagine, in only a few years' time the computer you are currently working with will no longer be available to you! The pace of technological change is rapid. The computer you'll be using in the near future will most likely be quite different than the one you have come to know so well. Think back to five years ago. During the last half decade alone, quite a bit has changed in the world of information technology. System architectures, chips, data carriers, operating systems, programs, and a whole lot more have come and gone--and it's a pretty safe bet that during the next five years still more changes will wash over us. Taking this line of thinking forward, what will the situation be like in 10, 15, or 20 years? More and more change awaits us. This realization leads us to ask again: can you recall which bibliographic references you assigned (all those years ago) to the object with the inventory number of

1014? Will your new computer be able to help you find the answer? The likelihood it will is low, quite low in fact!

Such a problem is easily overlooked, particularly since it will not even emerge for several years. That said, it must be considered quite seriously today. If we lose access to material in digital form, we also lose all the information we need to go about our work. All of our efforts to preserve collections for future generations could be at risk.

It's highly unlikely that many future programs will allow us to read old files, and it's equally unlikely that future operating systems will accommodate readers for older storage media. Everything surrounding the data we are creating is undergoing change. The only workable approach to the situation is to change our data as well. Our data need to be constantly adapted to new software programs, and they need to be transferred from one physical storage medium to the next.

Once completed, an index card can be placed in a cabinet. Twenty or even fifty years later, a colleague may retrieve the card and read all of the information you recorded. That's not at all how the digital world of data files works. Conservation of digital information requires ongoing attention. CDs, for example, should be copied in their entirety every 2 or 3 years; otherwise, you won't be able to access the data you've stored on them. Every time a new program, or a new version of a program, is introduced at your workplace, you should check to see if the data you've created in the past remain accessible. You may need to save them in a new file format. Electronic files will not be conserved unless you take action to preserve them.

When you create a digital file you are already affecting how much work will be required to preserve your information in the future. How your data should be preserved – and how the labor required to do so can be kept to a minimum – is the subject of this booklet. Our aim is to spell out as concretely as possible precisely what can and must be done.

Anyone who creates or bears any sort of responsibility for data files should take to heart the proposals outlined in the following pages. As only a few of us are computer specialists, we'll offer concise answers rather than elaborate technical explanations, examples rather than theorems, and, where necessary, to avoid technical jargon we'll provide more detailed explanations. After you've finished reading this booklet, you, too, should be in a position to declare, "Actually, it's all rather simple."

It's All How You Look at It

Anyone who's interested in conserving files should ask what happens when you hit "save." The question seems straightforward enough--trouble is, it's not the right question. We humans are not in fact "saving"; rather, we are instructing a computer software program to record information in the form of a file to a particular location on your computer's hard drive. We're not splitting hairs here. The form in which your information is put together in a file depends on the program you use to save it, though programs offer us a certain degree of latitude. We can save image, text, video, or audio in this or that data format – that is, as long as the software program we've chosen allows us to do so. Other programs use other data formats. And this often leads to misunderstandings.

To see what we mean, try using Adobe System's Photoshop program to save an image in the PSD data format. Now attempt to open this file using another program. Chances are you won't succeed. However, Adobe Photoshop would have allowed you to save the image in the TIF file format, and a broad range of different software programs would have enabled you to open a TIF. In other words, many different programs "understand" which information and what location you wish to have interpreted when you choose this format (but not PSD).

We'll have more to say about data formats later. For now, what you need to understand is that programs and data formats are closely connected. Most software programs are able to read or save only a limited number of file formats. Let's assume the range of data formats a program offers doesn't meet your requirements. You'd have to find different software. But what options do you have? You can't run every software program on any given operating system. Of course, you could replace your operating system as well, but here, too, your choices are not boundless. Not every computer is suited to run each and every operating system on the market. Anyway, we are free to choose the computer we want: provided, that is, we can afford a new one.

In the computer world, everything is interconnected, with one set of choices placing (often unforeseen) limits on another. When we ask a computer to save our information, we have already made a number of choices, both implicit and conscious: for a particular type of computer, operating system, software program, storage medium, and recording device. Most of the time, we don't really have to give much thought to these decisions, that is, until it comes time to ensure that our data will be available for years to come.

Look at it this way: anyone who stores data uses a particular software program to do so. This program operates with a particular operating system on a certain type of computer. With this program you are saving data in a particular data format on a particular storage

medium. To read from this particular storage medium, you need a specific type of reader. This reader must in turn be operable with a certain type of operating system.

To simplify matters, let's give this creature a name. Every piece of data has what we'll call its own formation environment, and the components of this formation environment are reciprocally dependent. While we're at it, let's also provide a name for the space in which files are accessed. Let's call this space the use environment. The use environment and formation environment share the same types of components. In both environments, the components are reciprocally dependent on one another

If we save a file in a particular environment and then immediately access it in the same environment, then the formation and use environments are identical. That's simple enough. If, on the other hand, we save the file and then attempt to access it via another computer, we are bringing the file into another environment altogether. Whether we'll be able to read the file depends on the extent to which the two environments differ. It's the same story when we save a file and then forward it to a colleague who'd like to open and edit it. Differences between the formation and use environments will affect whether the file in question can be used.

This view of two different environments illustrates that the chances of our being able to read our files in a few years hence depend on various factors. What's obvious is that storage media degrade over time. For instance, after only a few years many disks and CDs are no longer readable. What's less obvious is what frequently leads to headaches: the files may have been preserved on various storage media but are actually no longer useable because the programs we need to interpret them no longer exist. Problems also arise when the files have been preserved on a medium (a floppy disk, for example) for which there are no longer any readers. Furthermore, antiquated programs and readers often require long-forgotten operating systems, and so on...

Any difference introduced in the various components of both environments – use and formation – reduces the likelihood that we'll be able to access a given file in the future. Difference, or as seen across time, change, is at the root of the problem. Finding a solution might seem simple: all we need do is ensure that nothing else changes in the world of technology. Unfortunately, it's not as simple as that. Eventually, even the most robust computer will give up the ghost; someday, even the most durable CD or disk will refuse to access your precious data. And perhaps one day there won't be any programs to support the file formats we chose so carefully all those years ago. In other words, we can't avoid change. Should we then stand by idly as our data fade into the fog of history? Although we are in no position to prevent change, we can attempt to shape our response to it in constructive ways.

Make Changes with Care – and Control for the Effects

Any change to computer hardware and software presents a danger to the existence of your digital information. Anyone planning to buy a new software program should determine in advance whether the old files will be accessible via the new program. By the same token, anyone seeking to purchase a new computer should check to see if the software used to access storage media in the past will run on the new computer without any glitches.

To someone who has always saved to disks in the past, what good is a super-fast new computer that won't allow you to hook up a disk drive? Ditto a new operating system that won't accommodate magnetic tape readers if these tapes have served you well in the past.

Before you make changes to the environment in which you use files, it is important to determine whether the new environment you have chosen will permit full access to the information you have saved in the past. And after every change to a new environment, you will need to check to see if the old files remain fully accessible. You should avoid changes that appear to bring initial advantages at the cost of invalidating your old data.

Make Changes in Good Time

That said, there are many good reasons to embrace change. To limit ourselves to only one example, anyone planning to spend more time in the future working with video files will sooner or later want to part with her or his aged computer. The constant expansion of technology offers still further enticements. A more compelling reason to accept change may come when your old computer finally crashes. Given the challenges associated with long-term digital preservation, you don't want to wait too long to make needed changes.

A program saves files in a certain predetermined format. The structure of the file and the quantity and type of additional information it provides, also known as the format, may deviate from the format that was used in previous versions of the program. In other words, what frequently accompanies each new version of a program is a new version of the format. That's not always readily apparent, as the manufacturer of the program may retain the old format's label (such as ".doc"). In the short term, a vendor's decision to "update" the format does not seem to present any special problems. The new version of the program is by and large programmed to read files produced in the most recent version of the software. However, before too much time passes, serious problems may emerge. The introduction of a subsequent version of the program may reveal an inability to read files produced in the program only two versions earlier – files produced in still older versions are unknown to the program. Graphically, the problem can be summarized in the following way:

	File Format 1	File Format 2	File Format 3	File Format 4
Program Version 1	XXX			
Program Version 2	XXX	XXX		
Program Version 3		XXX	XXX	
Program Version 4			XXX	XXX

Readable Formats

Anyone who produces files with Program Version 1 – and then changes to Program Version 2 – will be able to read and files in the first version. Troubles begin when our user then changes to Program Version 3. Now she or he can read files created in Program Version 2, but not in Program Version 1.

The only practical course of action available is to load the initial files – the ones saved in Program Version 1 – with Version 2 of the software. Before closing the file, save the file in the format available in Version 2 of the software.

So far so good. The file format can be read in Program Version 3. However, when Version 4 comes along, it'd be advisable to bring the files created in Program Version 2 forward one version, in other words, to load and save in the new version of the software, and so on.

If we stick with this example, all of the files in question must be reloaded and saved anew each time a new version of the program comes along. Adopt this approach and you are going to have to stay on your toes; if you skip one or more updates of the program, you could find yourself unable to read your data! If you follow this leapfrog approach to preservation, you shouldn't wait too long to make changes; otherwise, you'll miss one of the versions of the program you need to keep your data accessible.

Perhaps you'll decide to change the program altogether, rather than to update constantly all of your older files. In that case, you should open all of your older files in the new program and save from it. This increases the likelihood that subsequent programs will be able to read your data (in the meantime, maintain a copy of older formats of the new program as well). In effect, changing programs requires more or less the same degree of time, energy, and diligence as version updates.

Each transfer to new formats poses risks. While new programs (or new versions of an older program) are often able to read files, they are often unable to fully interpret them. Footnotes may no longer end up exactly where they should be, or you might face unexpected challenges integrating graphic elements and images. With each new conversion, the likelihood that you will encounter serious differences between the original file and the most recent version increases – in other words, the chances that some of your information will be lost multiply along the way.

	File Format 1	File Format 2	File Format 3	File Format 4
Program Version 1	XXX			
Program Version 2	(XXX) →	YYY		
Program Version 3		(YYY) →	ZZZ	
Program Version 4			(ZZZ) →	AAA

Changes to the Contents of Files through Migration

The only way to avoid this valley of troubles is to never enter it – or to escape at your earliest opportunity! The good news is that accomplishing this is easier than you might imagine.

Minimizing the Effects of Change

The problems we face with changing formats stem from the fact that the power to shape and name them rests with the manufacturers of software programs. It is at their discretion, not ours, whether to make changes. Because the newest versions of their programs offer enhanced capabilities that are saved together with our files, the formats (and in a sense our files) are constantly under further development. Data formats controlled by software manufacturers are called proprietary file formats. Often enough, the definition of the formats they sell are not disclosed so as to prevent competition from other software developers.

The good news is that users do have options. To offer just one example, Microsoft Word employs proprietary data formats. Controlled and defined by a private interest, the program nonetheless also allows users to save files as "text only." The resulting files are very small; they do not contain additional information about fonts, font size, and font color. Those are the disadvantages of the "text only" option. On the plus side, you gain valuable storage space. More important, files saved in this format can be read by a variety of programs (and this is true of both older and new versions).

The ASCII format has been with us since the late 1960s. There is a high probability that files saved in ASCII will be accessible via a wide variety of computer programs well into the future. We say this for essentially two reasons: we possess a variety of documents in this format, and the program can be used by any developer without paying license fees. Whenever you feel you can do without special text symbols (as well as fancy presentation of footnotes, easy integration of graphics, and other such frills), you should save documents in the ASCII format (ASCII files often bear the ".txt" file extension name).

When you use proprietary data formats you are placing yourself at the mercy of others. What happens if the software company you are using disappears one fine day? The file format is either secret or subject to copyright protection. Data formats that belong to a particular supplier but whose definition has nonetheless been made available to other

companies for further development present a better option. So long as these "open definition" formats are actually incorporated in a variety of different programs, these types of data formats are actually relatively stable. Because many people working with many different programs will save files in this format, developers of future programs will be inclined to include this "open definition" format in their latest products. An example of this type of data format is the Tagged Image File Format, better known as TIFF or TIF. Formally under the control of Adobe Systems, the .tif definition has been disclosed and is available to all interested parties.

The best option is to use data formats that are widely used, disclosed to all users, and free of copyright restrictions. Examples of such formats include the .txt format for text mentioned above, the .jpg format for image files, and the .mpg format for video files; the latter two files, JPEG and MPEG, were created by an international consortium. The Moving Pictures Expert Group (MPEG) consists of 350 industry and university representatives. The Group's recommendations have been accepted by the International Standard Organization (ISO). The same is true of the work of the Joint Picture Experts Group, or JPEG; their specifications have also been raised to the level of international standards. Use of the JPEG format is open to all. For this reason, it's possible to create and combine JPEG files with a large number of programs.

We'll have more to say about individual formats. What's important to grasp here is that it is possible to minimize the adverse effects of change. The means to this end are simple: rather than using proprietary data formats, chose those nonproprietary formats that are open to all and used by many. The same thought process and conclusion apply to storage media. When a particular storage medium is widely used, when many software companies are manufacturing and selling machines that read and write on it, the likelihood is greater that both the storage medium and the hardware devices used to run it will be available for years to come. Special storage devices like the Zip drive may be of use to you in your day-to-day labors, but when it comes to the long-term maintenance of your digital files, we strongly advice against their use, as both the drives and media are produced by only a handful of manufacturers.

From the perspective of continuing access, and given the components of the formation environment as described earlier, you should always choose the generic over the specific. Extravagant solutions may seem to offer advantages at first blush, but over the long run they may turn out to be a hindrance. Sooner or later, you'll have to transfer your files to a more widespread data format or storage medium. The work and trouble you'll one day face can be reduced if you make conscious choices about appropriate data formats and storage media the first time you click "save."

For more information:
http://aida.jiscinvolve.org/toolkit

File Formats

As discussed in the previous section, you should avoid formats subject to the exclusive control of any one supplier. The same holds true for formats with undisclosed specifications or definitions. Instead, rely on more widely used formats. Before offering specific format recommendations, we'd like to discuss a few basic considerations you should keep in mind when it comes to the long-term preservation of your digital data.

Formats can be divided into proprietary and nonproprietary (sometimes called "open") formats. They can also be classified according to the degree of their dissemination, that is to say, whether they are widely dispersed (or less so). Another way to categorize them is according to their intended purpose. Text file types are fundamentally different than image file formats, and these in turn are decidedly different than video file formats. In other words, we cannot offer one general recommendation to cover all cases.

That said, all formats can in principle be divided according to whether or not the data they contain has been compressed. One might assume that for the safekeeping of data over long periods of time, compressed data formats would be preferable – after all, compression ensures that file sizes are considerably reduced, and certainly in the case of video formats, compression is a must. Without compression, e.g., the concentration of data according to a specific algorithm, the volume of video data would far exceed the capabilities of the average computer. So as far as video files are concerned, compression will remain part of our preservation future. But when the choice is, to evoke Shakespeare's Hamlet, "to compress, or not to compress," what decisions should users make?

As it happens, there's an important reason not to compress files that we want to be able to read many years from today:

Computers generally work in two states, reducing all information to binary opposites. That's how digital data is stored: via a tiny ridge (indicating "yes") or its absence (indicating "no") on a CD or DVD, or by directing small magnetic fields in a particular direction (indicating "yes") or by their absence (indicating "no") on either a diskette or a magnetic tape. In the same way, deep inside your computer only two possible conditions exist. Everything the computer takes in by way of information via the keyboard, microphone, camera, etc., is either translated into binary form or arrives as "binaries." In other words, once it enters the computer, your collections information becomes a jumble of ones and zeros. Absent a special device, the human eye cannot see digital data. Nor, for that matter, are they readily understandable. The information that you want to preserve has been encoded: thus, if you want to read your file, hear your song, or watch your movie again, the digital files in question must always first be decrypted.

In order for decryption, the reconstruction of your data, to succeed, it is imperative to know the specific code your computer has used to save your information. And what happens when we compress a file? The already-encrypted data is encrypted yet again. To gain access to your files, now we have to know two (or more) codes. The more encryption codes you have, the greater the likelihood that you'll end up losing one (or more) of them: and that's as good a reason as any not to compress files. There's absolutely no assurance that the programs used to carry out these encryptions will be around a few years from now, or that they'll run on the computer you'll be working with then.

This piece of advice, to avoid compression altogether, applies first and foremost to formats such as ZIP. In addition to these types of compressions, that – at least at the moment of their creation – create encrypted files that nonetheless contain all the relevant information you need, there are also compressions that work in such a way as to extract information from your files. An example of this type of "lossy" compression is the saving of image files in the JPEG format. Small, manageable, yet large enough for the computer screen, JPEG files also travel quickly over the Internet. These are clearly important advantages. And yet, conservation of .jpeg files is by no means advisable. As mentioned above, the creation of files of this sort involves nothing other than the destruction of part of your collections data. All the same, the format is widely used, and its compression algorithm has been disclosed, e.g., it can be used by any software manufacturer or Internet user. And if you're charged with the long-term preservation of a Web site loaded with JPEG files, you really don't have many options. However, if you want to conserve a photographic collection of museum objects in digital form, why would you choose to do without as much information as possible? If you want to manage digital information over time, it's clearly better to create a file that preserves all the available data. For those occasions when you need to use a particular file (for instance, when you want to include a photograph of a museum object on your museum's Web site or attach a copy in an e-mail to a colleague), you'll always be in a position to create a compressed file according to the quality and size needs of the moment – that is, as long as you have taken our advice to establish an image archive that does not consist of JPEG files.

In light of these considerations, the maxim to choose always the generic over the specific can be expanded in two interrelated ways. To ensure their availability for many years to come, data should be as complete as possible and saved in a form as simple and as widespread as possible. Data should thus be collected in a way that is nonproprietary, that is, without restrictions to anyone's use.

With these criteria in mind, we now turn to our discussion of the use of several common formats.

For more information:
http://www.gdfr.info

This global registry of digital formats was launched in April 2009. Project partners include in the US and UK national archives; the world library site, OCLC; Harvard University Library; and the Andrew W. Mellon Foundation.

Really long-term preservation (hieroglyphs from Luxor)

Formats in Detail

1. Image Formats

TIF

Description:
"TIF," or rather "TIFF," is the abbreviation for Tagged Image File Format. Aldus developed TIFF in cooperation with Microsoft and others. A sixth version of .tif is currently available. Originally owned by Aldus, copyright was transferred to Adobe when the two firms merged in September 1994. The format's original specification dates back to the 1980s. Version 6.0 differs from Version 5.0 in only one respect: it states that Adobe owns the definition. The last major change to this format definition took place in 1988 with the upgrade from Version 4.0 to Version 5.0.

The TIF format affords very high color depths. The size of the largest possible file is limited to two raised to the power thirty-two – a limit in the gigabyte range more theoretical than real. The TIF format can save multipage images or documents to a single TIF file (as opposed to a series of files for each scanned page). It's also possible to save files in other formats within the TIF format (for example, a JPEG within a TIF). Many programs allow one to compress the files – the common form is a Lempel-Ziv-Welch (LZW) data compression algorithm – of .tif data when you hit "save."

Control:
A single firm (Adobe) controls the definition, though not its use by others. That means there are no restrictions or license fee requirements. The format's definition is available to anyone. Any software manufacturer can use the open definition to create new programs. A conscious effort has been made to keep TIF flexible. Software vendors continue to develop their own TIF variants.

Distribution:
The format is widespread. The number of programs that can read and write .tif files is impressively large.

Pros and Cons:
The main disadvantage of the .tif format is the size of the resultant files. The chief advantage is that all information pertaining to an image file can be saved in a TIF. The number of programs that can work with TIF files is an additional advantage, not to mention TIF's utilization of a very large color space.

Assessment:

The TIF format is suitable for long-term preservation. That said, there are three main points to keep in mind. Avoid compressions, saving other file formats within TIF, and including more than one image in any single .tif file.

JPEG

Description:
The JPEG image coding standard does not refer to a format but rather to the organization that created it: the Joint Photographic Experts Group, or JPEG. The image format standard this group created is called JFIF, or JPEG Interchange Format (JIF). JPEG's (JFIF's) format's definition was last revised in 1992. In August 1990, the definition was declared a norm (ISO/IEC IS 10918) by the International Standards Organization (ISO). The International Telecommunications Union, the United Nations Specialized Agency in the field of telecommunications, declared JPEG Recommendation T.81. This so-called baseline standard allows only the creation of lossy compressed files. Later, the standard was expanded to permit users to save uncompressed files as well. This new standard is referred to as JPEG-LS Standard (ISO/IEC IS 14495-1 | ITU-T Recommendation T.87). Parts of this new standard are however subject to a patent held by Hewlett Packard. That said, use is explicitly open to anyone. This observation also applies to a further iteration of the standard according to which the data are saved in such a way as to allow one to reconstruct the image from the data in a series of steps, also known as the "Progressive JPEG" format.

High color depth can be saved in JPEG, and it's generally possible to select the degree of compression when you create a .jpeg file. The larger the compression, the more information you lose. When this happens, artifacts can emerge when your images appear on your computer screen. These square areas of colors – the colors are similar to those displayed in the picture – emerge within the displayed image.

Control:
The .jpeg format is supported by a large group of firms and universities working in cooperation with international organizations for standardization and communication. The baseline definition is free of patent restrictions; its use is not subject to restriction. Files created according to a later JPEG standard, JPEG-LS, or Lossless JPEG, are on the other hand subject to many patents. Nevertheless, interested parties are invited to use JPEG-LS' patented algorithms without restrictions.

Distribution:
The JPEG format is widely used on the Internet and as a storage format for digital cameras.

Pros and Cons:
The chief advantage of the .jpeg format is the small size of its files. The disadvantage lies in the lossy nature of storage.

Assessment:
In terms of long-term conservation, we can only recommend the .jpeg format to a limited degree. Although the wide distribution and disclosure of the definition speak for its recommendation, the fact that the only way to be sure you have avoided patent restrictions is by saving your image data in a loss-afflicted compression clearly speaks against it. By including JPEG files in your digital repository, you have decided to conserve incomplete files. Wherever possible, you should chose the TIF format over JPEG.

If – for whatever reason – the choice is nonetheless made in favor of .jpeg, avoid saving JPEGs in either the "progressive" format or in JPEG-LS. Both are subject to patents, and may at some point lead to licensing restrictions.

$$F(u,v) = \frac{\Lambda(u)\Lambda(v)}{4}\sum_{i=0}^{7}\sum_{j=0}^{7}\cos\frac{(2i+1)\cdot u\pi}{16}\cdot\cos\frac{(2j+1)\cdot v\pi}{16}\cdot f(i,j)$$

$$\Lambda(\xi) = \begin{cases} \frac{1}{\sqrt{2}} & for\ \xi = 0 \\ 1 & otherwise \end{cases}$$

Application of the discrete cosine transform (DCT) used to compress JPEGs

JPEG 2000

Description:
The disadvantages of the JPEG format (some of which are reviewed above) led the Joint Picture Experts Group in 2000 to develop a new format, JPEG 2000. In 2001, the format was published as an ISO/IEC 15444. Through the use of different compression algorithms, the new format avoids so-called JPEG artifacts. Image files of comparable quality are much smaller than in their original JPEG format. JPEG 2000 also enables you to save in a "lossless" mode, in other words, to save without losing valuable information about your images.

Control:
Some parts of the format definition are subject to patent declarations. For the most part, however, the format can be used without restrictions. To display JPEG 2000 files, you may have to change the settings of your browser, image editing programs, etc. You may also have to use special presentation software.

Distribution:
While it's true that a few larger cultural institutions, such as London's Wellcome Trust Library, are currently using JPEG 2000, the format has been slow to find adherents. Software manufacturers have not been quick to embrace JPEG 2000 files.

Pros and Cons:
JPEG 2000 is much better at compression than conventional JPEGs. In principle, the ability to compress files without loss of data makes this format interesting for continuing access well into the future.

Assessment:
The limited distribution of the format leads us to argue against the use of JPEG 2000.

GIF

Description:
In January 1995, Unisys decided to charge royalties for the use of its LZW (Lempel-Ziv-Welch) compression algorithm. With this decision, the firm assessed fees for the use of GIF, the Graphic Interchange Format, a format that has existed since the early days of the Internet. Compuserve and Unisys developed the format together; in 1995, Unisys held the patent. The firm also charged a fee for the compression of TIF files in instances where the compression was carried out on the basis of this particular algorithm. GIF format version 89a allowed the use of animated images. The US patent expired on June 20, 2003. Counterpart patents in France, Germany, Italy, and the United Kingdom expired on June 18, 2004. In reaction to Unisys' demands that they pay to use the GIF format, more and more software manufacturers turned to another format, PNG (Portable Network Graphics, see below). To ensure that older browsers are able to read their Internet Web sites, some programmers continue to use .gif.

The GIF format is limited to 256 colors. The format enables one to display transparent surfaces and to store more than one image in a single file.

Control:
Compuserve, later Unisys, held this format's definition. GIF's compression algorithm was patented. Patent restrictions might again apply one day. The dispute over GIF licensing fees is illustrative, as was programmers' decision to drop .gif in favor of other formats with similar properties.

Distribution:
The format remains widespread, in part because many older browsers remain able to interpret Web sites with .gif files.

Pros and Cons:
Beyond the patent dispute described above, GIF's chief disadvantage is its limited palette of 256 colors. On the other hand, GIFs are very small in size, presenting distinct advantages.

Assessment:
We recommend against saving image files in the .gif format, mostly because owners of the underlying algorithms have in the past demonstrated their willingness to exploit their market position. We believe software vendors will increasingly turn away from .gif. That means the likelihood you'll be able to find a program that'll enable you to manipulate GIF files will dwindle over time.

Additional Image Formats

From the multitude of remaining graphics formats on the market we'll restrict our attention to the best known. All of those listed below are unsuitable for the long-term maintenance of files. These formats are in the possession of a single software manufacturer. What's more, they can only be created, read, or edited from a small number of programs.

BMP (Microsoft: Bitmap)
CPT (Corel: PHOTO-PAINT file format)
DNG (Adobe: Digital Negative)
PNG (Portable Network Graphics)
PSD (Adobe: Photoshop file format)
PSP (Corel: Paint Shop Pro file format)
UFO (Ulead: PhotoImpact file format)

Avoid these formats when it comes to the long-term conservation of your image and graphics files. They may disappear sooner than you think.

2. Text Formats

DOC

Description:
Microsoft owns the DOC format. It is very widely distributed. The DOC format saves in a single file text together such markup elements as bold and italics. DOC files may also contain macros or graphics. The DOC format is highly complex and closely intertwined with the Windows operating system. In the past, when Microsoft has redesigned Word, its flagship word processing program, the company has done so without ensuring DOC format compatibility. Strictly speaking, for this reason you can't really speak of one .doc, but rather multiple DOC formats, some of which are unfortunately incompatible with one another.

Files saved in an older version of DOC formats, i.e., files saved in an older version of Word, cannot be opened in newer versions of the program without considerable difficulty. At present, there are seven versions of DOC in circulation. Microsoft introduces a new version of DOC every two years or so. Even with the latest, ostensibly "XML" version of the DOC format, DOCX, Microsoft continues to march exclusively to the sound of its own drummer, rather than to adhere to standards agreed upon by the industry as a whole. With DOCX's arrival, the time has come to ensure your textual materials are safely stored in another format altogether.

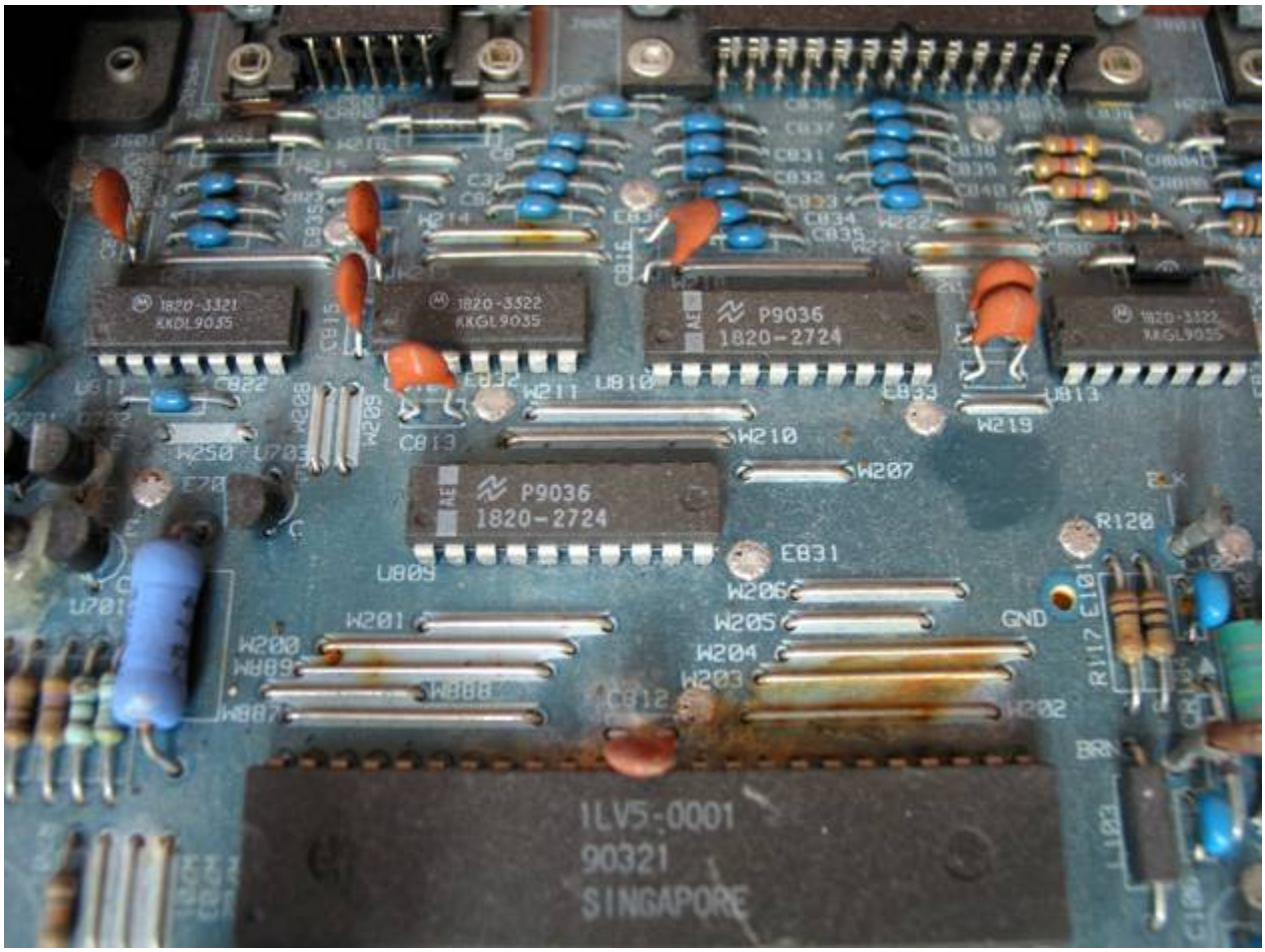
Control:
Control of the .doc format definition rests entirely with the Microsoft Corporation. The definition has only partially been disclosed. Although Microsoft will provide the definition upon request, the firm then demands that requestors maintain strict secrecy.

Distribution:
The format is very widespread.

Pros and Cons:
Even when you attempt to save text in a single font type and size stripped of special markups, the .doc files you're left with are still quite large.

Assessment:
The readability of files saved in the DOC format is not assured even today, much less in the distant future. Even if the recipient of a .doc file is a running a version of Microsoft Word, she or he may have trouble using the file if they're using a different version of the program than the sender. If the recipient of the .doc file is running a different word processing system than the one in which the file was created (in this instance, Microsoft

Word), then the contents of the file in question may well be entirely inaccessible. For the purposes of preserving long-term intellectual access, therefore, this format is clearly inadvisable: the troubles we are having today sharing textual information via .doc are a harbinger of things to come.



Past its expiration date. An artifact found on the side of the road.

RTF

Description:
In response to the compatibility difficulties created by its DOC format – problems that have extended across not only different versions of Word and other software programs, but also the Windows operating system itself – Microsoft introduced the RTF format. With this change, Microsoft's word processing software was configured with the option to save documents as .rtf files.

RTF (Rich Text Format) files can be read in other programs. That's the good news. Unfortunately, however, Microsoft changes the definition of RTF with almost every new version of Word. Still, before we run down the list of negatives, we should note that there are a few good points to RTF. First, Microsoft is paying attention to .rtf version compatibility. At least equally important, the .rtf format is independent of any one particular operating system (such as Windows). Files saved in the RTF format are based on ASCII text (see the next section on TXT) interspersed with control orders.

Control:
Microsoft maintains exclusive control of the .rtf format definition, but at least the specification has been disclosed. And we should add that Microsoft has encouraged other software companies to design their programs so that they can read RTF files as well. That said, Microsoft could withdraw its support for RTF at any moment, and it should be remembered that the company could also simply refuse to further develop the format at some point in the future. Alternatively, Microsoft could choose to make RTF incompatible with other proprietary programs it owns or may one day develop.

Assessment:
Although it's more suitable for the long-term preservation of text files than .doc, effective control of the RTF format rests with only one technology company. For this reason, we cannot recommend use of the .rtf format.

TXT

Description:
TXT actually stands for the ASCII (or ANSI) format. In this format, text strings are stored in a single row at a time. Markup elements are not included – for instance, text marked bold is saved, but without the bold marking. Neither cross references nor inline graphics may be included in this format. On the positive side, this format is compatible with virtually any word processing program: reading and writing TXT files presents almost no difficulties whatsoever. Files extensions for .txt vary. Sometimes, files may be saved as TXT, in other cases as ASC. Still other programs offer the option to save in "Plain Text." All yield the same result. ASCII is the basis for HTML, XML, and the RTF format.

ASCII stands for the American Standard Code for Information Interchange. ASCII files have been in use since 1963. Modifications were necessary not long after its introduction, as the initial version failed to take account of non-English characters. In response, in 1968 the American National Standards Institute (ANSI) developed the ANSI code. Thereafter, letters not found in English were assigned a unique number equivalent. Microsoft Word has decided to label the format TXT: to save files as TXT within the program, choose "Text Only" in the drop-down menu under the "Save As" command.

Control:
Strictly speaking, ASCII/ANSI code represents a mere translation table available to anyone. The format consists of strings, nothing more. Neither the arrangement of the text strings nor the use of the table is patented. No restrictions apply to use.

Distribution:
As the basis for other formats, ASCII/ANSI is widely disseminated. Nearly every word processing program allows one to save and read in the ASCII format.

Pros and Cons:
A major disadvantage of the format is its inability to save text markup elements (such as cursive) or other elements often associated with text documents (such as graphics). A decisive advantage of the format is its compatibility across all operating systems and with almost every other computer program.

Assessment:
ASCII/ANSI code has been around since the earliest days of computation. Whenever you can live without text layout, use it.

PDF

Description:
PDF stands for Portable Document Format. The format's definition is the property of Adobe Systems. The format was created as an elaboration of the page description language Postscript. Together with Adobe Acrobat, a program created to display files in this format, .pdf was introduced in 1993. The definition has been disclosed; as of July 1, 2008, PDF has been published as an open standard (ISO/IEC 32000-1:2008). Adobe provides a free program (Acrobat software) for those who wish to read and customize .pdf files. With the aid of free additional programs, Internet browsers are able to display PDF files.

Control:
Adobe maintains exclusive control over the .pdf file format. In contrast to Microsoft's approach to the DOC format, however, Adobe has disclosed PDF's definition; it is available to any interested party. Nonetheless, the format remains proprietary.

Distribution:
PDF is a widely used and accepted document standard. In a sense, .pdf is on its way to becoming the electronic equivalent of paper. Free read-only programs, coupled with its small file size in relation to the complexity of the textual content, have made .pdf very popular.

Pros and Cons:
The PDF format was created as a page description language for printers. Understanding how the individual textual elements are structured in .pdf files requires extensive technical knowledge. This complicates the search for information within and between files, as well as the transformation of .pdf content into other formats.

Assessment:
Because of its widespread dissemination the PDF format is, albeit only to a limited degree, suitable for the purposes of long-term archiving. Problems begin with the fact that the PDF is controlled by only one vendor (admittedly, this manufacturer has made the definition available to all interested parties). PDF is a highly complicated format, a product of its design as an aid to printers.

PDF/A

Description:
PDF/A was established in response to the difficulties associated with preserving long-term access to textual materials. It owes its existence to the success of PDF, a file format, as discussed above, that has since the early 1990s enjoyed considerable success in smoothing the flow of text from desktop computers to printers. PDF/A comes in different shapes and sizes: the two you'll hear the most about are PDF/A-1a (sometimes referred to as PDF 1.4) and PDF/A-1b.

Control:
This effort to create a standard format for the long-term archiving of electronic documents dates back to October 2002. As it did with PDF, Adobe Systems has shared the definition of PDF/A, allowing others to engage in the open development of new standards. The initial impetus to establish PDF/A came from a number of major software firms, together with the US National Archives and Library of Congress. Interestingly, the Library of Congress initially hesitated to recommend PDF/A to its employees. As it happens, not all documents that appear to be saved as PDF/A are in fact true to the PDF/A format standard.

Distribution:
Dissemination of the PDF/A file format among companies and larger not-for-profit organizations is growing.

Pros and Cons:
One of the PDF/A file formats, PDF/A-1, has been an ISO standard (ISO 19005-1) since 2005. Much of the information necessary to display your document in the same manner today as in the future (text, fonts, colors, etc.) is included in the PDF/A-1 file. Unfortunately, however, the PDF/A-1 standard is a remarkably tough read. Aside from the technical points, at more than a thousand pages, its volume alone is quite daunting. And although it's less than five years old, a new version of the ISO is already on its way.

Assessment:
While simplification of PDF/A's ISO is a welcome development, the fact that the various parties involved in establishing PDF/A almost immediately began to draft a replacement is not a good sign. Any documents you choose to save in PDF/A-1 will, it's true, most likely be viewable in future computer environments. The look and feel of your document in the future will thus be preserved. As one tool among many, PDF/A has its merits. What it's not is a one-size-fits-all solution for preserving electronic documents.

For more information:

This 2007 study by the Dutch National Library hints at some of the difficulties posed by PDF/A:
http://www.kb.nl/hrd/dd/dd_links_en_publicaties/PDF_Guidelines.pdf

3. Video Formats

To ensure you are able to enjoy a video without any noticeable judders or jerks, at least 25 images must be displayed each second. The resultant file sizes are immense. As we explained above, the compression of your data – and all the work they represent! – should be ruled out as a matter of principle. For video, however, we have to make exceptions: all those fast-moving pictures lead to files too large to manage without compression, and might require further encryption of your video data collection.

Given the fact that compression appears unavoidable for the near future, it's essential that you choose a standard that is both widely distributed and supported by a variety of manufacturers. The standard must also be available to others so that the files you save can always be reconstructed in the future.

The Motion Picture Experts Group (MPEG) has developed several such standards since its establishment in 1988. MPEG is a working group of the International Standards Organization (ISO), in which many leading manufacturers of software and hardware are represented. While Microsoft offers the AVI (Audio Video Interleave) format, and Apple MOV (also known as QuickTime), the MPEG format of the Motion Picture Experts Group is independent of operating systems and manufacturers.
At present, the current MPEG standards are as follows:

MPEG1 for video CD and mp3 files
MPEG2 for DVD and digital television
MPEG4 for multimedia applications
MPEG7 for analysis and search in videos.

The specifications of the MPEG group are limited to the use of particular algorithms. Software manufacturers can, for their part, integrate additional compression algorithms in their file formats.

We recommend that you pay very close attention to the MPEG group's recommendations. If at all possible, make sure that any additional file format definitions you're using are not controlled by any one firm alone.

Examples of proprietary video formats include:

AVI Audio Video Interleaved (Microsoft)
FLV/SWF Flash Video (Adobe Systems, originally Macromedia)
MOV Apple Quicktime
WMV Windows Media (Microsoft)

4. Audio Formats

Similar to their video cousins, sound reproduction files have not yet undergone full standardization. The standard established by the Motion Picture Experts Group (MPEG) in 1987, mp3, continues to gain significant ground on its competitors, but mp3 has not yet carried the day.

The WAV (or WAVE) format is a file format developed by Microsoft and IBM. For the most part, it works without compression. As a preservation medium, it's at best suitable for relatively small audio collections.

MIDI (Musical Instrument Digital Interface, sometimes called MID) does not allow users to create high-quality music files. On the plus side, .midi files are very small.

Apple's AIFF format is not widely distributed.

While we are currently unable to offer a clear recommendation for sound archiving, we'd like to point to a recommendation offered by the International Association of Sound and Audiovisual Archives (IASA). In response to the IASA's efforts, museums and other cultural repositories have adopted PCM WAV 96 khz/24 bit as a common standard.

For more information:

This August 2006 study from the UK's Arts and Humanities Data Service provides an overview of sound and moving images archiving:
http://ahds.ac.uk/about/projects/archiving-studies/moving-images-sound-archiving-final.pdf

The following EU portal is a good source of information on audio and video preservation:
http://www.tape-online.net

The site below provides a case study on selecting audio formats:
http://www.arl.org/bm~doc/soundsavingstableofcontents.pdf

5. Database Files

Databases are composed of many different parts. Database management systems enable us to create small programs that yield forms, queries, reports, and the like. The resultant data is typically organized in the form of tables. Individual files (for administration, programs, and data) are saved in one or several additional files. Many different file formats are used to store database files.

In contrast to image or text files, which by and large contain something in its entirety, databases are dynamically designed. In other words, their capacity to assimilate and retrieve information is perpetually open, so they can take in more and more information over time. Thus, for the long-term archiving of digitized information, when it comes to databases we can speak only of saving snapshots in time. This observation applies not merely to the programs running in a particular database (that is to say, what's generating our forms, our queries), but also to the data itself (such as output tables). Generally speaking, users are not in a position to influence the format in which the smaller programs running in a particular database are saved. The format used to save output tables, on the other hand, is something many database management systems allow you to choose.

If your database system will allow it, from time to time you should export your data in the CSV (comma-separate values) file format. In this format, individual entries are stored in rows of TXT (see above for more details on .txt) separated by commas.

6. Formats, by JHOVE! AONSer is here!

What is a format anyway? In technical terms, a format consists of rules that allow you to map between your content (text, images, video, audio, and the like) and the bit streams your computer understands. If the aim is to preserve knowledge and context, and not just exceedingly long strings of 1s and 0s, we need to step back briefly from our discussion of individual formats to consider how you're going to manage all the various formats likely to pour into your digital repository in the years to come.

JHOVE

Two tools are currently under development to assist you in this effort. The first is JHOVE (pronounced "jove"). The JHOVE2 project is funded by the US Library of Congress as part of the Library's National Information Infrastructure Preservation Program. JHOVE seeks to answer two questions. The first is: "I have a digital object, what format is it?" Once that's settled, you're ready for the second question: "I have an object purportedly in format X; is it in fact format X? JHOVE is a potentially important tool for complex formats, among them, PDF/A. The future may well belong to digital objects that mix multiple files and formats in ways we now may find difficult to imagine. For that reason alone, it's worth visiting the project's Web site from time to time to consider the group's most current recommendations:

http://confluence.ucop.edu/display/JHOVE2Info/Home

AONS

The Automated Obsolescence Notification System (AONS) informs you when file formats are obsolete or at risk of becoming obsolete. The project is run by the National Library of Australia and the Australian Partnership for Sustainable Repositories. A beta version of the software is available:

http://sourceforge.net/projects/aons/

For more information:

Administered by the UK National Archives, the PRONOM project also offers highly useful information on file formats:

http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Storage Media

The challenges associated with the long-term preservation of digital data are bound up with the use of suitable storage media. We have intentionally placed the subject of storage media after our consideration of file formats. One might assume that the life of information saved electronically is identical to the life span of the storage medium in question. In practice, however, the data storage medium often remains intact, whereas the files saved to a particular data carrier have become inaccessible owing to the fact that suitable programs are no longer available to us. Still, the question of which data storage medium is most appropriate for long-term preservation remains important.

The storage media available to us at the present may be divided into two general categories: magnetic storage media, such as magnetic tapes or hard disk drives, and optical storage media, such as CDs, DVDs, and BDs (Blu-ray disks). There are still other types of storage media; because their dissemination remains limited, we won't cover them in this edition. As mentioned above, when it comes to the preservation of digital assets, the generic is infinitely preferable to the specific. Unless they're widely distributed, sophisticated storage media and the writing and reading devices that go with them can quickly disappear from the market, dashing all of your efforts to ensure your data lasts well into the future. We recommend generic solutions for two additional reasons. One is financial. Technologies and products that are widely disseminated have become so only because they are offered at a reasonable price. The second involves data exchange in the here and now. The likelihood that the recipient of your data will be in a position to interpret them, i.e., that he or she will possess the necessary hardware and software, is much greater if you avoid high-end data storage solutions. These may provide the more elegant and perhaps even most effective way to store information, but in the end, they are also most likely to become obsolescent.

Wide distribution is one key criterion you need to assess when considering storage media. Thinking forward in time, what's critical is not only widespread distribution of the storage media, but also the general availability of reading and writing devices--the hardware, as it were--that you will need to interpret the information you have saved electronically. There are still other criteria to take into consideration:

- Media must have a long life span.
- Media must be robust.
- Storing media should not place unreasonable requirements on users.
- Media should possess sufficient capacity to enable us to store files without compression.
- The hardware in question, i.e., the viewing and writing devices, should be easy to learn to use.

- Media should be structured in a way that makes it easy for users to reach the files they need.
- Media should be priced as inexpensively as possible.

Before we move on to our discussion of the different types of storage media, we wish to offer a more general comment. Given how technology works, it is very unlikely that there will ever be a single, perfect medium, "the one," "the ultimate," "the definitive" storage solution. The technology industry does not work toward refining or perfecting a single product: instead, one form of technology usually just replaces previous ones. Think about the ways we used to save audio. Shellac and vinyl records replaced wax cylinders. These, in turn, gave way to tape cassettes. Today, we still have CDs; but for how much longer? With every change in medium, music lovers are required to shell out more money to hear their favorite tunes. And while some, but by no means all, will hold out until their storied old horn gramophones fail, even such loyalists must eventually give in. The same will be true for users of digital data. Industry's need to turn quarterly profits means that corporations will continue to produce a steady stream of new storage media and shiny new instruments necessary to interpret them.

Environmental conditions affect all storage media. Pick any storage medium you like: after a certain period of time, you'll still have to copy your information to another medium of the same type (refreshing) or to a medium of a different type altogether (migration). How long any given medium can defy the laws of nature must by necessity remain vague. No one can say for certain. Manufacturers, for their part, tend to speak of long, sometimes very long, periods of time. In the case of CDs, we're told they may last for as long as 200 years. Should we take these claims at face value?

These figures bear little relation to office realities--how could they, given that the media haven't even been around that long! So that they can issue statements about the durability or life span of their products, manufacturers subject their media to so-called aging tests. In these tests, media are exposed to high temperatures and high humidity over a certain period of time. Media life spans are then calculated on the basis of test data. In 2002, an entirely new norm, ISO 18921:2002, was created for tests estimating the life expectancy of writeable CDs.

The first CDs were introduced by Phillips and Sony in 1980, and since then we've been told, time and again, that they'll last for 10, 50, maybe even 100 years. To date, no supplier has been required to adhere to the standards of the above-mentioned ISO; in other words, no one has replicated exactly the test conditions set out in the international norm and informed us precisely how long CDs will ostensibly remain useable. In practice, the information vendors provide us about the estimated life expectancy of their storage media is based on in-house tests carried out under conditions defined by the firms in question. Results, perhaps needless to say, are interpreted according to in-house criteria. But even if

vendors were to adhere strictly to this ISO, you have to ask yourself how useful a test that limits itself to two factors, humidity and temperature, can be. Many other factors also affect the life span of CDs. It matters how often a medium is read, or whether an environment is free of dust and other environmental contaminants. The amount of sulfur or other chemicals in the ambient air matters, as does the degree to which the medium is exposed to ultraviolet radiation. The list doesn't end there. For all of these reasons (and still others), it makes sense to adopt a highly critical attitude toward these so-called product tests and the claims some firms choose to divine from them.

What we do know for certain is that the life span of our storage, that is to say, the period during which we can be certain to gain access to the information we have saved, depends on how conscientiously each of us approaches storage and handling.

In the following sections we will assess the degree to which diskettes, ZIP drives, magnetic tapes, hard drives, CDs, DVDs, BDs (Blu-ray discs), and memory sticks may serve as building blocks in our efforts to maintain digital information across multiple generations of computer technology. We will consider each individual media according to such criteria as dissemination, durability, capacity, ease and speed of access, and price. For each media type, we will then offer our advice as to how you should maintain storage media so that they remain useable for as long as possible.

Media …

Diskettes

Description:
Diskettes, sometimes called disks or floppy disks, were once widely used. In recent years, disk drives have disappeared from new computers. The edge length of the first floppy disks was 8 inches (roughly 200 mm). Since then, disks have shrunk from 5 ¼ (roughly 133 mm) to 3 ½ inches (90 mm). This storage medium consists of a data disk encased in a plastic shell (hence the name "disk"). A magnetic layer has been applied to the disk. Within this layer, data are stored according to the specific alignment of magnetic fields. Although diskettes have grown smaller and smaller over the years, the amount of data that can be saved to a disk has increased exponentially as a result of technological advance. That's because data have been placed closer and closer together over time, making it possible to reduce the size of the magnetic fields. A tiny mechanical arm glides across the magnetic fields, reading or writing data; this arm touches the surfaces of the thin disk. Hidden from view within its case, the disk spins at something like 300 revolutions per minute. The disk's index is located at a predefined location (sector 0).

Distribution:
Although they have disappeared from many but by no means all offices, diskettes will mostly likely be with us in one form or another for some time to come. Floppy disks were long a popular means to exchange information (generally the upper limit for disks is 1.44 megabytes, or MB) between computers not connected to one another via a network. USB sticks (see below) have hastened the demise of the disk; ditto the advent of digital photography, with its need for storage media able to accommodate larger file sizes. Although hand-to-hand data transfer has indeed moved on, many an office cabinet contains stack of disks, a fact anyone establishing or maintaining digital files ignores at his or her peril.

Handling and Storage:
While we cannot recommend diskettes as a means to ensure that your digital memory will be available in the future, we can offer a few suggestions about how to store diskettes so that they will remain readable for as long as possible:

- Place disks in a light- and dust-free environment.
- Keep disks a safe distance from magnetic fields: do not place them near loudspeakers, older monitors, etc.
- Maintain a storage room temperature of 20 degrees Celsius.
- Ensure that the storage room's humidity is between 40 and 45 percent.

- Never touch the inner storage disk – the one that contains your data – with your fingers.
- Do not use a hard pen or pencil to label the medium; the pressure from your hand could destroy the data disk inside the case.

So long as you strictly monitor storage conditions, disks should, according to some manufacturers, last as long as ten years. In fact, this is anything but a cut-and-dry rule. You should know that every time you attempt to read old files, or create new ones, you are reducing the disk's life expectancy. If you use the disks quite often, don't expect them to last longer than five years; to be on the safe side, you should be copying data from diskettes to new storage media annually.

Assessment:
The history of the diskette is a tale of many shapes and sizes. It's also a story of rapidly changing – and rapidly disappearing – hardware. Because disks have fallen out of use, they are clearly unsuitable for long-term retention. If you have important data saved on disks of any kind, the time has come to move your files as quickly as possible to another storage medium.

ZIP Disks

Description:
Much of what we had to say about floppy disks (see above) applies to ZIP disks. Although ZIP disks were designed to offer considerably more storage than diskettes, they were unable to drive their floppy cousins out of the market altogether. Not long after their introduction, yet another medium, the user-writeable CD, came on the market. At that time, the CDs' storage capacity was still greater than that of ZIP disks. Another strike against the ZIP disk was its price. Licenses and fees for the ZIP technology belonged to one company, Iomega. The absence of competition drove up the price, limiting the medium's dissemination. Our recommendations for the storage of ZIP disks are analogous to those offered for diskettes (see above).

Assessment:
Just like other diskettes, ZIP disks are unsuitable for long-term archiving. Using ZIP disks means you are at the mercy of a firm that can at any given point in time cease production of the ZIP disk and its attendant hardware, such as the read-and-write head.

Magnetic Tape

Description:
Disks, as discussed above, store data on a magnetic disk. So-called magnetic tape is also a medium for (you guessed it!) magnetic recording. These tapes consist of a thin magnetizable coating applied to a long strip of plastic. This long strip of plastic is wound on to two reels.

There are essentially two different ways to arrange data on magnetic types. One is succession in a long loop. The other is set at angle to one another.

The alignment of data exerts considerable influence on the speed and frequency with which one can roll the tapes backward or forward. These actions apply pressure to the reeled tape. So also does the application of the read head. The time period within which stored data can be read without losses depends largely on how often you access information stored on the tapes. Magnetic tapes have been in use since the 1950s. Since the mid-1980s, tapes are no longer stored in open coils, but rather in so-called cartridges. Transfer of data from an old magnetic tape to a new one can be performed in automated fashion, at predefined intervals by machines that are (sometimes) called tape robots. This allows for at least the partial automation of long-term digital conservation efforts.

Distribution:
In many organizations, data backup – in the sense of a security copy – occurs via magnetic tapes. Tapes are often used when large amounts of data need to be conserved.

Handling and Storage:

- Insofar as possible, place tapes in a dust-free environment.
- Corrosive gases and chemicals damage magnetic tape.
- Tape rolls or cartridges that have been in storage should not be placed in the tape drive until after they have been given time to adjust to new climatic conditions.
- Relative humidity should be kept below 60 percent, and room temperature should remain between 18-20° Celsius.

- The rolls or cartridges should be kept upright and never placed on top of one another.

- Dropping the rolls or cartridges can negatively affect the uniformity of the coiling, leading to read error and tearing.

- Touching the tape surface can easily lead to data loss.

- If magnetic tapes are not rewound over a long period of time, you may encounter a "push through" problem. What happens is that magnetization at one point in the roll appears somewhere else, such as in another part of the coil where you first noticed the "push through" problem.

- Do not store cartridges in either paper or cardboard boxes. Special plastic cases do a much better job protecting your tapes from dust particles. These special storage cases must shut tightly so as to hold the cartridges firmly in place.

- Depending on how often you read (or write to) magnetic tapes, an exchange of media is from time to time essential. To be on the safe side, you should swap out the old media for new at least once a year, even if the tapes or cartridges have hardly been used.

Assessment:
Tape storage capacity continues to increase by leaps and bounds, with data being placed closer and closer together. Historically, tape has enjoyed a cost advantage over disk storage; whether this trend will continue is hard to know. For now, securing data by magnetic tape through the use of tape robots remains the standard method of long-term preservation at many institutions and firms.

Automated copying initially appears to be a huge plus, given the impressive data capacity tape offers. And yet one major problem remains: your files still need to be transferred periodically to new formats. Magnetic tapes come with other, more serious drawbacks. One is the long period of time required to access information. Another is that the medium's life span hinges upon how often you access it.

Magnetic tapes are a poor fit for the growing number of us requiring frequent access to electronic files. They are also not much of a solution if you can't afford a tape robot: without the automation such a machine allows, considerable time and effort are required to copy magnetic tapes. Acquiring new media at short intervals is costly, as is the amount of time you or other staff members would need to invest to retrieve valuable pieces of information. Perhaps the right solution for some long-term archives, magnetic tape is unfortunately not a good fit for many others.

Hard Drives

Description:
Hard disk drives (usually shortened to hard drives) are constructed in much the same way as floppy disks. Just like diskettes, hard drives magnetize many layers of material to represent binary digits, either a 0 or a 1. Unlike diskettes, hard drives combine and seal the read-and-write-heads in one casing to ensure that dust will not interfere with the workings of the computer. Sealed together, the read-and-write heads operate in nanometer-range (a nanometer is a billionth of a meter) closeness to the magnetic surface.

Despite the proximity of the heads and the magnetized materials, the two do not actually touch. Saving materials to a hard drive is thus free of both friction and contact, which enables you to access your files much more quickly than you could by diskette or magnetic tape. Hard drives are incredible spinning machines. A central component of the hard drive, the hard disk platter, turns no less than several thousand times per minute. The mechanical stress on the material is considerable. And the way hard drives are built, with all these incredibly small parts crammed together, means they are very sensitive to bumps or other types of movement.

Over the course of the hard drive's lifetime (exactly how long it'll last must remain a matter of conjecture), lubricants can evaporate to the point that the hard disk drive no longer plays. More and more compact, they're capable of taking in and processing larger and larger amounts of information. While typically a sealed unit you never actually see, there are some so-called removable hard disks (also known as external hard drives) that allow you to transfer impressive amounts of data between non-networked computers – just as we did, albeit on a smaller scale, in the halcyon days of the humble diskette.

Handling and Storage:
- Hard drives are susceptible to damage by bumps and other types of movement.
- Hard drives can be vulnerable to magnetic fields.
- Before using removable hard drives, allow time for the device to acclimate to different levels of humidity and other environmental conditions.

Assessment:
Hard drives combine high storage capacity with quick access. With the price of hard drives falling, it's tempting to ask whether an archive of hard drives is not a viable answer to the challenges of digital conservation. Whether they will serve as an ideal vehicle for long-term preservation, however, remains an open question. For one thing, they remain quite sensitive to movement; one slip of the hand and all of your work may well be lost forever. Magnetic fields and even changes in altitude present other challenges. Whether hard drives are a real alternative to, say, magnetic tapes remains to be seen. On the plus

side, they allow easy and rapid access to collections information. Time will tell whether the price of hard drives will fall to a point that magnetic tape manufacturers began to lose ground in the marketplace.

Gradual data loss (as we used to know it)

CDs

Description:
CDs have been with us since the 1980s. Originally developed by Sony and Phillips, today both CDs and the hardware players used to play them are manufactured by many different companies. As regards the long-term archiving of information, only two types of CDs, recordable compact disks (CD-Rs) and rewritable compact disks, (CD-RW), are of interest to us. Each CD (by 2007, over 200 billion CDs had been sold worldwide) is constructed according to the same pattern. On the underside of each CD is a protective layer; generally made of polycarbonate, this layer is transparent. Over it lies another layer that bears your data. Different manufacturers use various materials for this part of the CD. Directly above this data carrier layer is a lacquer coating that reflects the light of the laser. A fourth layer is also protective. Sometimes printed, this layer closes the CD from above.

Light penetrates the material from below. Adding data generates heat, changing the surface of the layer that contains your data. In the process, you end up adding a series of microscopic bumps to your CD. When data are read, light from the laser is either scattered through inconsistencies in the data carrier level, or it reaches the lacquer layer and is simply reflected. Data are reconstructed on the basis of incredibly small bumps and valleys that the laser is able to interpret as light signals. Differences among CDs emerge from the fact that different materials are used for both the reflector layer, such as gold, silver, and aluminum, and the data carrier layer, such as cyanin and azon. The respective combination of materials used to create these two layers determines the color of the CD's underside. Various materials and mixture ratios are used by different manufacturers. For rewriteable compact disks (CD-RW), different materials are also used in the data carrier layer. Audio CDs have a different logical structure than data CDs (CD-ROM). CDs created in an industrial pressing plant have still different qualities. Unlike burned CDs, they do not feature elevations and depressions. Instead, data is added to the disk in one physical stamping operation.

The construction principle behind the CD is in all instances similar. Data on a CD consists of a single spiral track stretching as far as six kilometers.

It's hard to say what's better, shimmering gold or silver luminous CDs. We do know that the manufacturers keep changing the composition of the materials they use to create the data carrier layer, and that some CD reading devices appear better suited to interpret certain carrier layers than they do others. While they have the same basic composition, rewriteable compact disks, CD-RWs, are slightly more complicated. For instance, they possess an extra recording layer both below and above the data carrier layer. And the layer where your data rest is made of different materials (common mixtures include silver, indium, and antimony). In contrast to the "pits and lands" written onto a polycarbonate

surface, a CD-RW disk boasts miniscule changes in the crystal structure of the recording layer itself.

CDs hold up to 700 MB of data. That means many millions of elevations and depressions must be placed on a disk with a 12 cm diameter. The individual wells in question are a mere 0.9 millionth of a meter in length. The device used to read and write CDs must work within incredibly tight margins of error. This incredible density of data comes at a price: anything that alters the precise course of the laser's light, such as scratches, deflections, dirt, or fingerprints, is either interpreted incorrectly or not at all. For this reason, CD drives are equipped with error-correction mechanisms that allow the device to reconstruct missing or falsely interpreted information on the basis of a saved checksum (a technique computers use to ensure the validity of data). But even a deeper-than-average scratch or a visible fingerprint is enough to throw a wrench in the works.

On the plus side, the CD's correction programs don't affect performance. Compared to magnetic tapes, CD technology allows fast access to stored data. What's more, CDs enjoy a clear price advantage over most competing storage media products.

Distribution:
CDs in all their various forms are very widely disseminated. The hardware and software needed to read and write CDs are available for all major operating systems.

Handling and Storage:
Given the precision with which the technology works, CDs seem at first glance to be impressively robust. In practice, however, some real differences among CDs emerge. In general, pressed CDs are much more robust than burned ones. And CD-Rs are much sturdier than CD-RWs. After only a short period of use, CD-RWs are no longer completely readable. Most of us lack the technical and financial infrastructure to produce our own pressed CDs. The following suggestions regarding the handling and storage of CDs apply to CDs you've burned yourself.

Much, but by no means all, of the following advice can be found on the packaging data of the CDs you purchase.

- CDs should be kept away from direct sunlight, which can lead to chemical alterations.

- Extreme heat and high humidity can lead to varying degrees of expansion within a CD's individual layers, resulting in a curvature of the upper surface. Chemical reactions can be set in motion. Moisture can make its way into the individual layers from the CD's outer edge, leading to unevenness. The best storage conditions consist of a relative humidity between 20 and 50 % and a room temperature at or below 20 degrees Celsius.

- Rapid changes in temperature or humidity should be avoided as they lead the materials that compose the CD to become slightly brittle.

- Store CDs in a vertical position in specially designed plastic containers. Avoid placing CDs on top of one another, as deformations can occur through the pressure.

- Fingerprints, especially on the CD's underside, lead to read errors.

- Do not use ballpoint (or many other) pens to mark the CD, as they may produce pits in the reflector layer, leading the laser to commit errors. Water-based, solvent-free pens are much more suitable; use only these.

- We advise against placing labels on your CDs. They're seldom centered properly, which leads to rotation imbalances and thus additional stress on the read-and-write heads. The adhesive materials may also influence the disk's upper protective layer. Chemical reactions may result, allowing substances to penetrate the reflector layer. Finally, parts of the label may come lose, causing more significant hardware problems.

- Add descriptions only to your CD's transparent inner ring. Check before you start writing whether the area allotted is sufficient.

- To reduce environmental degradation, do not unpack CDs to which you intend to add a description until you're actually ready to do so. Before you add your description, be sure to check the CD's surface (especially the underside). Sometimes you can notice trouble spots, such as a scratch, with the naked eye.

- CDs should be removed from the disk drive immediately after you're finished with them. Your computer and drive offer different climatic conditions than the plastic case you should use to store CDs.

- Check the readability of your recordable CD immediately after you have finished a writing session. If possible, use another drive to read the files than the one you used to write data. Open at least some of the files you've just completed, not just one.

- Adding new data to an already-initialized CD can lead to errors. In general, a so-called quad-speed drive (many modern drives are in fact much faster) is appropriate.

- Don't use rewritable media such as CD-RW. They are substantially more vulnerable to data loss than "write once" CD-Rs.

- Whether you're able to read data depends on the interplay between your recording device and the CD-R. Sometimes hardware manufacturers recommend specific CD-Rs

for use with their equipment. If no advice has been given, try out different types of CDs. Record and attempt to access files with different types of hardware.

- Don't despair if your hardware is unable to read one of your CD-Rs. It doesn't necessary mean that your data are lost. Try a newer piece of hardware, as it may have better built-in error correction capabilities. If you're able to access your files with the newer piece of hardware, by all means burn a new CD as a backup.

- Choose cleaning products that do not chemically react with the materials that make up your CD. Hard cloth or a coarse brush may lead to scratches large enough to destroy your data.

- Clean your CDs from inside to outside, or outside to inside, but never in concentric form. If you clean your CD this way, you'll be following the track of your data and run the risk of adding scratches that may wipe out entire blocks of data.

- Never attempt to blow on your CD to clean it. Your breath always conveys barely visible amounts saliva that may be deposited on the CD, affecting its readability.

- Read errors may also result from the contamination of the laser's lens. Before you decide the data on a given CD are lost for good, clean the lens. There are number of CD lens cleaning kits on the market; or, you can let a specialist handle it for you.

- Copy your data at minimum every 2 or 3 years to a new CD or another storage medium.

Assessment:
Experience has shown that CDs you burn yourself are far less durable than what manufacturers promise. Still, they hold up reasonably well in comparison to other storage media. Like other storage media we've discussed, their longevity is greatly dependent on environmental influences. While CDs are relatively inexpensive, portable, and widely disseminated, in the rapidly changing environment of information technology their days are numbered. Many new computers, such as netbooks and other laptop computers, are no longer equipped with CD drives. Others are outfitted with DVD drives that allow one to read and write CDs. Trouble is, CDs operate at a lower speed than DVDs. With a CD, your files are not packed as tightly on the disk as with a DVD. From the industry's perspective, CD technology is acting like a brake on how the DVD drive operates. As users demand DVDs that run at higher speeds, the future of the "slow" CD looks increasingly uncertain. It's not too difficult to imagine a future in which it will be hard to find a computer that'll allow you to access files saved on a CD. For the time being, anyone buying a new DVD drive should be sure to purchase one that plays as many CD formats as possible.

DVDs

The digital versatile disk or digital video disk (DVD) shares much in common with the CD, such as a common diameter and a circular shape. Similarities do not end there. Both CDs and DVDs consist of layers. Whereas all CDs have only a single data carrier layer, DVDs boast as many as four in which lasers add pits. The laser also works somewhat differently than the one used for CDs. Through the light of different wavelengths, the laser used to create DVDs adds far more information to an equally tiny amount of space. The pits are smaller and lie closer together than they do on CDs. As mentioned at the end of the last section, the DVD drive spins at a higher speed than does the CD's. The higher density of data and the fact that some DVD manufacturers dispense with protective coatings in order to save costs mean that DVDs pose some disadvantages. DVDs are more subject to scratches and damage from ultraviolet light. As with CDs, there remains uncertainty about which DVD standard will ultimately carry the day. Some manufacturers tout the virtues of DVD-R and DVD-RW, whereas others argue that another standard, DVD+R and DVD+RW, will prevail. Lacking a clear industry standard, DVDs remain at best a questionable medium to ensure continuing access; if you choose the "false" standard, i.e., you burn DVDs in a format that does not ultimately emerge as the industry standard, sooner or later you'll encounter difficulties accessing your information. Other problems also remain unresolved. One form of DVD, the DVD-9, sometimes also referred to as PowerDVD, is a single-sided dual layer disk. These disks continue to present compatibility problems for many DVD media players.

Although DVDs have been with us for fifteen years, DVD technology continues to undergo rapid changes, making them unsuitable for long-term data archiving. If you nonetheless decide to use DVDs to this end, you should in handling and storing DVDs follow the recommendations we offered for CDs.

Blu-ray Discs

Blu-ray discs (or Blu-ray, sometimes also BD) are the latest members of the optical disk family. Created to store high definition video and gaming data, Blu-ray disks have the same physical dimensions as the standard 12 cm CD/DVD disk. The first rewritable Blu-ray disk (BD-RE) dates back to 2006. The name of the disk comes from the "blue" laser (the color's actually violet) used to read and write data. Developed by the Blu-ray Disc Association, a group of companies that introduced the format in February 2002, the number of titles available in the new format remains limited in some parts of the world. Still, we expect Blu-ray disks to continue to grow in popularity in the coming years.

While certainly impressive in terms of data storage (Blu-ray discs permit you to save more than six times as much information as you can to a DVD), when it comes to preservation, Blu-ray is not a viable medium. We expect that Blu-ray will in a few years be surpassed by a host of cheaper successors, chief among them hard drives themselves, as the wheel of technical obsolescence, one of the few constants in the computer industry, continues to turn. The problems with information density associated with DVDs and CDs also apply to Blu-ray; in fact, they may well present more significant challenges. In sum, we regard Blu-ray discs as an unsuitable data archiving medium. If for whatever reason you decide to use Blu-ray to build a data repository, be sure to handle and store BDs according to the recommendation we offered above for CDs.

Integrated Circuits

As our discussion of CDs, DVDs, and BDs makes obvious, we can reasonably expect the storage medium of the future to be, in a word, smaller. At the heart of this trend toward miniaturization is a device as humble as it is ubiquitous. It is the integrated circuit, better known simply as the chip.
Chips constitute the most important components of the electrical and electronic objects we encounter in our daily lives. Integrated circuits have been with us since the mid-twentieth century, when scientists determined they could be used to perform functions previously carried out by vacuum tubes (vacuum tubes were critical in the operation of a new electromagnetic method to detect objects developed during the Second World War, known to us today simply as RADAR, or radio detection and ranging).

To create an integrated circuit, you need to place a series of minute structures in a bed of silicon. Today, a thin slice of extremely pure silicon, a wafer, is used to fabricate integrated circuits. This wafer passes through a series of highly refined processing techniques: taken together, these allow for the creation of millions of microscopic circuit elements on the surface of tiny chips of silicon. Some of the most detailed steps in semiconductor device fabrication are undertaken by what's known as a wafer scanner. The lithographic patterning performed on the wafer requires it to be properly positioned and brought into focus, time and again, through a repetitive process of "step-and-repeat," or "stepper" for short. In order to place as much circuitry on each wafer as possible, a minute choreography of automated positioning and synchronization must be repeated many, many times over.

The current industry standard for these structures is 45 nanometers (nm). This 45 nm technology "node" demands extremely precise positioning with a high repeat accuracy in the range of, a drum roll please, plus or minus 2 nanometers!

USB Sticks

Every hour of the day, wafer stampers churn out amazingly tiny integrated circuits for products all around us, including but by no means limited to USB sticks. Sometimes referred to as USB flash drives, these storage media are, like their optical cousins (CD, DVD, BD), removable and rewriteable. The first USB flash drives came on the market in 2000. The USB's brain, a small integrated circuit encased in plastic, is relatively durable, and its price and size make it popular in offices where disk drives have all but disappeared. That said, the integrated circuit within the USB stick has a major Achilles' heel: the number of write and erase cycles is limited. Exactly how many times you can count on the integrated circuit to deliver your information as requested remains unknown (though, as with CDs, DVDs, and Blu-ray disks, this hasn't stopped manufacturers from guessing).

USB devices are inexpensive. And they're handy, so much so that it's at times difficult to remember where you last left yours. Ease of transmission, the USB stick's forte, also comes with downsides. One is malicious software (malware) and other bugs designed to wreck havoc on an otherwise secure desktop or network. The fact that they're so small (and thus easily misplaced), coupled with the more serious drawback of the as-yet unknown number of read-and-write cycles available to users makes USB sticks unsuitable for long-term digital archiving.

Summary: Storage Media

So now we've taken a closer look at several different types of storage media. The list could go on to include a much larger number of second and third cousins--we could even extend the family tree further. However, everything we have covered thus far tells us one thing: as far as the question of long-term preservation is concerned, the choice of data storage medium is not the crucial issue.

There's no storage medium to which we can safely entrust our data for the next 100 years. All of the currently available media demand that we periodically copy data to a new carrier of the same type or to another media type altogether. Our choice of long-term storage media determines how often we'll have to make these new copies. After 5 years at the latest, the data you want to keep must be transferred to new storage media. That's true no matter what storage medium you've chosen. Data saved on magnetic tapes have to be transferred to new storage media more often than CDs. Your decision in favor of a particular storage medium affects the amount of time and work you'll need to invest in the conservation of your data. Financial considerations also play a role in such decisions, though differences in price among the various media storage types are not pronounced.

Our consideration of the various media types underscores the need for vigilance. Technological developments are constantly yielding new types of media. In the process, older types of storage media are crowded out, and as a consequence, the hardware we need to access data saved on older media disappears from the market. Given this reality, you have to be flexible. It may just be necessary to create a copy of your data on another media type long before you had intended to do so. Keep in mind that the hardware to access your files can disappear from the market fairly quickly. One way to reduce this risk is to always save your data at least twice on at least two different types of media.

The storage media currently on the market compel us to copy, time and again. All of us need to pay much closer attention to the intervals we set for our copying operations. To this purpose, you will most certainly need to mark all storage media intended for long-term preservation with a date. If at all possible, you should also establish a protocol that clearly defines when the next copying operation will be necessary. Finally, someone within your organization should be appointed to monitor storage and to ensure that copying of media actually takes place on agreed-upon dates.

You won't be able to guarantee long-term conservation unless you appoint someone responsible to handle the task and then define the steps you expect this person to carry out on specific dates. Data curation, not an end-all commercial repository stocked with the latest media, is the answer to the ongoing challenges of change over time.

You Can't Save Everything – Nor Should You Try

In case you had any doubts, a quick look into your e-mail box will probably convince you that not all of your data should live on ad infinitum. As you know from your work with physical collections, planned retention is much preferable to a "keep everything" approach. In the digital world, nonselective archiving dramatically increases storage costs, not to mention the amount of time you or a colleague will need to pinpoint missing information in the distant future. It may also leave you open to unforeseen legal challenges; many countries have enacted legislation to cover the archiving of digital materials. Laws governing personal data protection, Freedom of Information (FOI), intellectual property, design, copyright, patent, and database rights may apply.

Take a hard look at the digitized information within your institution and decide which files are only of short-term interest. Keeping in mind the general guidelines we set out in the first pages of this booklet, draw up a list of the types of information you feel it may be useful to save for a longer period of time. Define selection criteria in light of your organization's mission, legal requirements, and technical capabilities. Working closely with your IT department or vendor, evaluate how your data may be reused in the years to come. Share your plans with colleagues, especially from other organizations facing similar challenges. Delete files of short-term interest. Structure your long-term archive to include a limited number of formats and storage media.

These are just a few of the steps you'll need to take as you begin to prune the vines. And just like a vineyard, data archiving will require your attention on an ongoing basis. Digital stewardship is about much more than hardware, software, and content; it also encompasses a broader range of policies, workflows, budgets, services, and, of course, people. No commercial "e-repository" provider can provide you with all of the answers; you, the cultural heritage professional, are in the best position to say which long-term management strategy makes the most sense for your organization's collections.

For more information:
http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/appraisal-and-selection.pdf

This is a good introduction to the issues surrounding digital appraisal and selection.
http://www.life.ac.uk

This tool helps you to calculate all of the costs associated with digital preservation.

Further Considerations and Recommendations

Whenever you alter the environment in which you produce data, whenever you change your operating system, set up a new computer, or start using a new program, you should ask yourself whether all of your data will remain fully accessible. You may have to take action in any number of instances. Perhaps your new computer doesn't include a drive for your storage media. Or maybe your new computer does have a drive, but it doesn't allow you to access information on the CDs you burned on your old computer. Or maybe a new software program won't allow you to open older files. The list doesn't end there. In any event, whenever something changes in the environment you use to create information, you should ensure that your files will remain accessible. As long as your old computer still functions, you may be able to use it to recover your information and transfer your files to a new environment without any losses.

You should double-check to confirm access each time you copy a file. Always try to read at least some of the data you have just copied. If possible, use another computer to do so.

As a matter of principle, make several copies of the data you want to have available to you for many years to come. Save archived files to several different storage media. Store the media you have selected for long-term preservation at different geographic locations. Make sure the physical storage conditions are appropriate. Placing media in a variety of locations also helps to reduce the risks of data loss posed by natural catastrophes, fire, or theft.

Stay generic. Use software that is broadly disseminated, that is to say, common operating systems and storage media. Save your data in a format that's as widely distributed as possible, one that can be interpreted by the broadest possible number of programs.

Choose storage formats and media with care, keeping in mind your long-term preservation needs.

Don't get rid of still functional computers and programs right away; it may turn out that you'll need them again.

But most important of all:

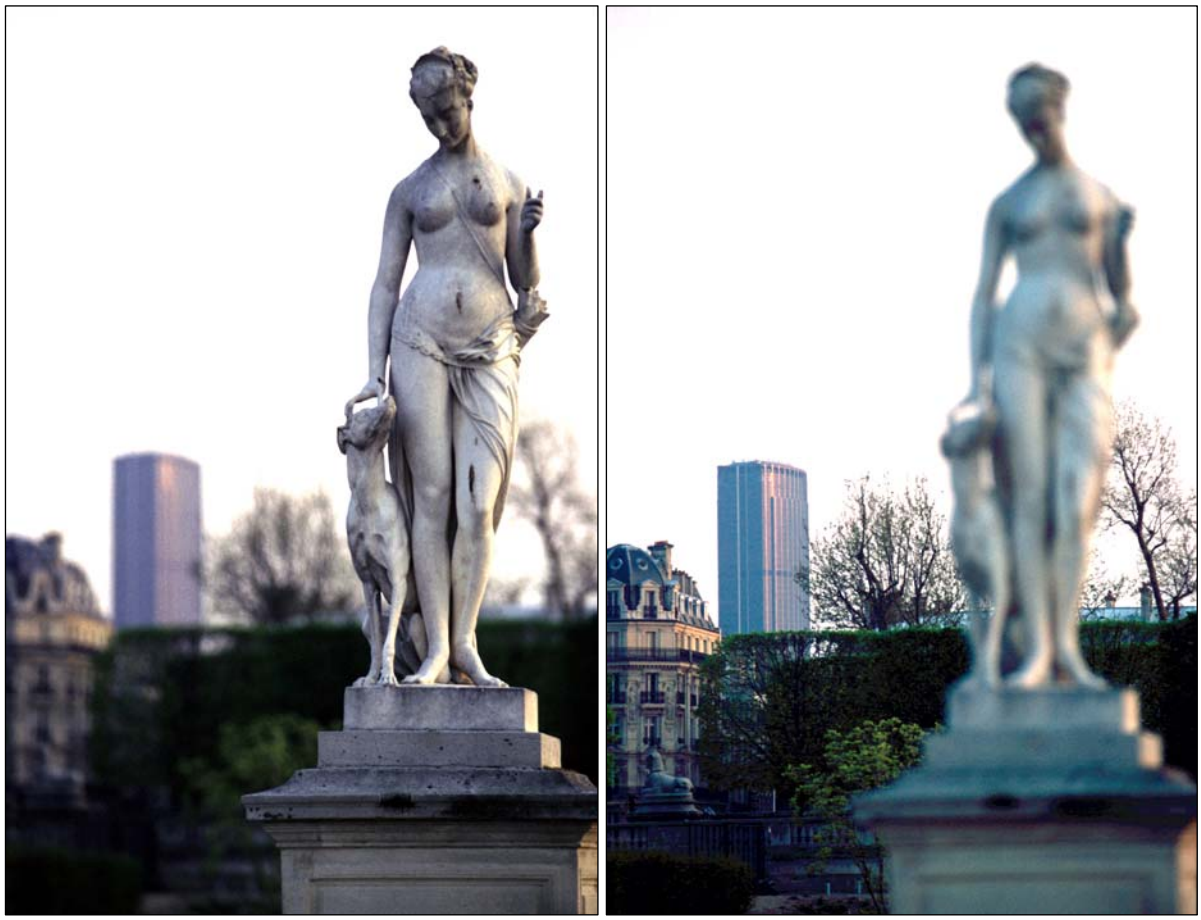
Take leave of the notion that to "save" files on your computer is identical to having them available to you for years to come. Consider the long-term conservation of your data as a discrete set of tasks that cannot be managed in a single effort: regard your stewardship of digital objects as a process you will need to manage actively. Be sure to designate those responsible for the task of conservation and develop a preservation plan that includes

steps, goals, and the means to monitor progress. Once you've taken these steps, there's a good chance you and your colleagues will be able to access and use your collections information many years hence.

See:

A Framework of Guidance for Building Good Digital Collections (available at the Web site listed below as a PDF):
http://framework.niso.org

Good Practice Guide for Developers of Cultural Heritage Web Services (available at the Web site listed below as an interactive document):
http://www.ukoln.ac.uk/interop-focus/gpg/



The past stands before the future.

Appendix

Overview of our recommendations

Formats

Image files	- TIF
Text files	- ASC, PDF/A
Database files	- convert to ASC
Video files	- no recommendation
Audio files	- no recommendation

Storage Media

Floppy disks	- unsuitable
ZIP disks	- unsuitable
Magnet tape	- limited suitability
Hard drives	- limited suitability
CD	- limited suitability
DVD	- limited suitability
Blu-ray	- unsuitable
USB sticks	- unsuitable

XML

As you learn more about the long-term archiving of digital data, you're going to come across XML. The abbreviation stands for Extensible Markup Language. XML's strength lies in its ability to do what you do best: describe collections. And providing lots of description, as any good museum hand will tell you, is the best way to ensure that a collection is kept alive for others to interpret and enjoy.

So what is XML exactly, and how does it work? To start with, XML is a computer language that is in many ways similar to another computer language you have probably heard of, HTML. In fact, XML looks a lot like HTML: XML files are saved as simple text (ASCII), just like in HTML. The key difference between them is that while HTML is a language to create Web pages, XML allows you to create a special language for your unique museum collection.

Another key difference between HTML and XML is that XML gives you a chance to add structure to your files, which allows data to draw on the strengths of your computer while remaining readable to humans. The key to this structure is the use of tags to identify data that can be reused by your computer or other computers. Hundreds of languages have been specially developed from XML for museums, libraries, archives, and other cultural heritage organizations. A boon for those seeking to share information about collections, XML makes use of a so-called schema to define the tags and other bits, called elements, attributes, and values, just like in HTML, you use to create your tailor-made XML application.

A set of markups has been defined for the use of XML files in museum databases. These markups have found considerable international support. Below is an example of an XML file in the Dublin Core, a XML application used by many museums:

<?xml version="1.0" ?>
<dc-record>
<type>Physical object</type>
<type>original</type>
<type>cultural</type>
<format>L:81 W:52 H:2</format>
<title>Kite, insect, fly ? </title>
<description>Kite, L:81 W:52 H:2, collected in China by Berthold Laufer (1874-1934) in 1903, depicting an insect, possibly a fly. Materials: paper, bamboo, pigment, string, masking tape (modern), metal (modern). Native term: Ts'ang Yin [Shang-Yin?].</description>

<subject>Kites</subject>
<subject>Insect</subject>
<subject>Fly</subject>
<contributor>Laufer, Berthold</contributor>
<publisher>American Museum of Natural History Division of Anthropology</publisher>
<date>1903</date>
<identifier>AMNH-ANT 70/10596</identifier>
<relation>HasFormat CD269/CD269/70/10596.PCD</relation>
<relation>HasFormat http://anthro.amnh.org/images/full/70/10596.jpg</relation>
<relation>HasFormat http://anthro.amnh.org/images/preview/70/10596.jpg</relation>
<relation>HasFormat http://anthro.amnh.org/images/thumbnails/70/10596.jpg</relation>
<relation>HasFormat AMNH-LIB 7683</relation>
<relation>HasFormat AMNH-LIB 338999</relation>
<rights>American Museum of Natural History Division of Anthropology</rights>
</dc-record>

[source: http://www.cimi.org/public_docs/meta_bestprac_v1_1_210400.pdf]

Object descriptions of this sort using standardized markups can be easily and automatically integrated in databases: in your own and those of others, whether they are down the street or halfway around the world.

The manifold possibilities of XML, its simple presentation allowing the representation of complex structures, and its growing recognition as an international standard make XML a good way to structure text files in keeping with the goals set out in this handbook.

Archiving E-Mail

Policies

As regards e-mail and other digital media, international experts view archiving not as a distinct intellectual endeavor, but rather as merely one component of a larger process of what's generally described as digital curation. For this reason, recommendations concerning the preservation of e-mail and other digital media begin with data creators, not record-keepers. In this view, those charged with the preservation of electronic resources act as advocates to define opportunities to shape not only the way digital media are stored, but also how they are created and used – both today and well into the future. In the process, depositors become stakeholders in the work of the archive, as functions once regarded as largely distinct – the creation, preservation, and access of documents – are brought closer together to one another.

The initial aim should be to standardize use of e-mail via an institution-wide policy that specifies basic archiving and retention procedures. The aim is to shape outgoing e-mails so that they conform insofar as possible to an expected pattern. Current best practice recommendations can be summarized as follows: customize e-mail headers and signature blocks to add contextual metadata. Ensure storage of official messages to IMAP rather than the POP3 message protocol to ensure valuable documents are not stranded on individual e-mail clients or personal home accounts. Within the archive, limit access to stored, e.g., in the possession of the repository, messages. Establish audit trails within the archive to trace actions by restricted and identified individuals to ensure no message is purposely compromised.

Legal Aspects

An oft-cited reason to adopt a comprehensive approach to e-mail stewardship is legal. To establish an archive's potential liability, each repository should develop a risk framework based on an information compliance audit. The key issue is the division of private from business e-mails. In addition to a range of Freedom of Information (FOI) and data protection laws relatively well known within the archival community, a wide range of intellectual property legislation, such as trademarks, patent, copyright, design, and database rights (renewable in many cases each time a database changes), apply to certain documents.

A review of these and other relevant directives should guide your archive's policy with regard to retention periods and the separation of personal and unofficial messages from official organizational documents. E-mails destroyed for compliance or other reasons must include copies of messages stored on back-up tapes kept for business continuity purposes.

Though viable for attachments and small text collections, PDF/A is not suitable as an e-mail archiving medium in light of the sheer volume of electronic messages likely to enter archival collections in the coming decade.

Technical Approaches and System Recommendations

Only a small percentage of e-mails will be valuable for historical inquiry, and as with legal aspects, the key issue in current professional discussions is authenticity. Information managers agree that the authenticity of e-mails is best achieved through preservation of as much as metadata as possible. They also stress that backup tapes lack the structure to explain message context, relationships, or facilitate retrieval. While there's a good deal of enthusiasm for XML-based approaches based on the oft-cited Open Archival Information System (OAIS) reference model, no suite of tools currently allows one to cover the entire OAIS. At the ingest end of the OAIS spectrum, a good deal of the current discussion revolves around authenticity requirements, e.g., which attributes must be kept to ensure an e-mail is regarded in the future as authentic. The trick is to capture as much as of the metadata as possible as soon as possible, that is, at the point of its creation. As regards e-mail, the best non-proprietary solution at present appears to be XENA, a free software created by the National Archives of Australia that detects file formats and converts them to open formats using extensible mark-up language, or XML.

Tempting as it may be, the solution is probably not to spend money on a turn-key system. Choosing a commercial vendor at this point in the development of e-mail archiving entails the significant risk of creating a system that isolates e-mail content and contacts, erasing the network features that define e-mail and its connections to other media. If the provider of a proprietary digital storage system goes out of business, so might your archive. Many commercial vendors favor PDF/A as a textual storage medium for e-mail. With reference to readability, accessibility, and management, a Belgian study published in 2006 came to the remarkable conclusion that proprietary e-mail storage systems offer no real advantages over off-the-shelf e-mail programs! In light of the archive's overarching goal to ensure continuing access decades into the future, the aim should be to implement open source software that allows you to release your content from the many legal restrictions of proprietary systems. Once an institution-wide policy toward e-mail management has been put in place, your interim solution should initially focus on storage of the original e-mail transmission file, RFC 2822. These files contain the metadata necessary to establish that the message was sent from the purported sender to the purported recipient and identifies the date and times it passed through each server. Log file of e-mail traffic should also be preserved in this interim digital depot to confirm which messages were sent, by whom, and when. RFC 2822 encodes attachments. The key is to decode attachments before they enter archival storage, converting these files insofar as possible to non-proprietary formats. Once this has been completed, each attachment can be relinked to the appropriate RFC 2822 files. (The fact you need to maintain links between RFC 2822 files and an

impressively long list of attached media also argues against purchasing and maintaining a document management system exclusively for e-mail.) RFC 2822 metadata also provide a good basis to manage the legal rights identified upon completion of your legal risk audit (see previous section).

Once these interim steps have been completed, test the best open source XML solutions. At present, the one with the strongest track record is from the National Archives of Australia. A recent promising addition is the so-called "parser" developed by the Rockefeller Archive Center and the Smithsonian Institution Archives. The purpose of this computer program is to migrate groups of e-mails into an XML file that captures e-mail records in situ, complete with their attachments, in other words, in the organizational context in which they were kept by the e-mail account owner.

When testing potential systems for the long-term e-mail repository, consideration of how e-mail content streams will be "re-used" by future researchers should inform discussion of access restrictions and future message display options. These needs will guide your choice of descriptive metadata, the counterpart to the preservation and administrative metadata obtained through RFC 2822 files. This stage will also entail a legal evaluation of access to collections via standard terms and conditions. Upon completion, the archive will require users to click an "I agree" box to obtain data. Here you can also place a link to a longer text outlining your repository's legal policy.

Select Publications:

Filip Boudrez (2006), "Filing and Archiving E-Mail," retrieved 4 January 2010 from http://www.expertisecentrumdavid.be/docs/emailrapport_lr.pdf

Maureen Pennock, (July 2006), "Curating E-Mails: A Life-Cycle Approach to the Management and Preservation of E-mail Messages," DCC Digital Curation Manual, S. Ross, M. Day (eds.)
http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-emails

Karin Schwarz (June 2009), "E-Mail-Archiverung," nestor-Handbuch, edited by Heike Neuroth, et. Al.
http://nestor.sub.uni-goettingen.de/handbuch/

"Developing a Policy for Managing E-Mail," (2004), The National Archives (ed.)
http://www.nationalarchives.gov.uk/documents/managing_emails.pdf

Pilot Projects:

Dutch National Archives e-mail preservation testbed, XmaiL:
http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=273

National Archives of Australia's XENA e-mail project:
http://xena.sourceforge.net/

Rockefeller Archive Center and Smithsonian Institution Archives' e-mail preservation parser:
http://siarchives.si.edu/cerp/parserdownload.htm

Additional Resources:

The National Library of Australia's resource collection on e-mail archiving:
http://www.nla.gov.au/padi/topics/47.html

RFC 2822 Internet Message Format:
http://www.faqs.org/rfcs/rfc8222.html

The US National Archives July 2008 guide to e-mail archiving:
http://www.archives.gov/records-mgmt/bulletins/2008/2008-05.html

The University College London and British Library's tool to calculate all costs associated with digital preservation:
http://www.life.ac.uk

Further questions? Consult Preservation Experts

Relevant information, including expert reports, legal advice, and information on upcoming workshops and training sessions, are available at these Web sites:

PADI
National Library of Australia, Preserving Access to Digital Information (PADI):
http://www.nla.gov.au/padi/

Nestor
"Nestor" is the Network of Expertise in Long-Term Storage of Digital Resources. Nestor's project partners are the:

German National Library
Göttingen State and University Library
Bavarian State Library
Humboldt University
Institute for Museum Research, SMB-PK
Library Service Center (BSZ) Baden-Württemberg
Landesarchiv Baden-Württemberg
University of Hagen, Department of Computer Science

www.langzeitarchivierung.de.

NDIIP
National Digital Information Infrastructure Program, US Library of Congress:
http://www.digitalpreservation.gov

DPC
The Digital Preservation Coalition (DPC) was established in 2001 to promote joint action on long-term conservation of digital information across the United Kingdom.
www.dpconline.org

Comments or Questions? Please write:

Dr. Stefan Rohde-Enslin
s.rohde-enslin@smb.spk-berlin.de

or

Dr. Keith R. Allen
kra@keithrallen.com