Conference “Imatge I Recerca”
Girona, Spain
November 16-19 2004

Longevity of digital raster images

Author:
René van Horik (rene.van.horik@niwi.knaw.nl)

NIWI-KNAW
Netherlands Institute for Scientific Information Services
P.O. Box 95110
1090 HC Amsterdam
The Netherlands

Abstract

The main subject of this paper concerns the longevity of a specific type of digital object, namely a digital image. This paper consists of three parts.

The first part of this paper contains background information on digital images. An explanation is given on the fact that there is a large number of digital image formats. Also the difference and relevance of digital born images and digital images that act as surrogates of an analogue original is covered in the first part of the paper.

The second part of this paper gives an overview on the current state of art in preservation of digital objects. The main digital preservation strategies are explained and the emerging OAIS reference model is covered. This international standard is an important conceptual framework for a wide range of digital preservation initiatives. Other topics discussed in the second part of this paper are preservation metadata, file format preservation and the durability of storage media for digital objects.

The third part of this paper concentrates on the preservation of digital raster images. The main factors that determine the longevity of a digital raster file are the applied image file format and the application of preservation metadata. Detailed information on the most suitable image file format and preservation metadata with regard to the long-term access of the digital images is given. If used in a specific way (e.g. without compression) the TIFF image file format seems to be the most durable digital raster file format. A number of metadata element sets do exist that improve the longevity of digital raster images. Some metadata element sets aimed at the formulation of technical metadata for digital images are discussed. The availability of technical metadata is an important facilitator for the longevity of digital raster images.

Introduction

The problem of the creation of stable and permanent images or pictures was not solved with the invention of photography in 1839. This is illustrated by the following quote[1]:

> *"... the daguerreotype image was as fragile as a butterfly's wing, fleeting and much more difficult to reproduce than an engraving. There was a general consensus that photography would become a force only once it could produce durable, infinitely repeatable images. ...This ambition had been partially achieved by the end of the nineteenth century, but did not reach its full commercial maturity until later."*

By the end of the twentieth century image capture and reproduction by means of digital devices became available on a wide scale. And again the question of the stability and permanence of images, now in digital form, was put forward, illustrated by the following quote[2]:

> *"Digitisation of cultural artefacts should provide a lasting electronic record for scholarly and universal access, preservation, and study. At the present time, however, digitisation projects are proceeding without established methods of recording precise conditions of digitisation."*

Obsolescence of the digital image format and deterioration of the digital data storage medium are among the main factors that threaten the long-term access to digital images. A number of digital preservation approaches exist to prevent that digital objects are not accessible anymore in the future. Based on the current state of art, this paper reports on what can be done to minimise the risk that digital raster images created today cannot be rendered anymore in the future.

The first part of this paper discusses general background information on digital images. The next section of this paper describes general digital preservation strategies as they are commonly accepted at the moment. The function of this section is to place the digital preservation of digital objects in a broader context.

The third part of this paper concentrates on the durability of digital raster images. The main issues that determine the long-term access of digital raster images are the file format used to arrange the digital file that contains the pixels and the quality of the documentation or metadata.

1. Digital images

The computer processing of images began in the USA at the National Bureau of Standards in 1956[3]. Several decisions made by the developers of the first scanner have influenced engineering practice ever since, e.g. the usage of rectangular arrays of square pixels. A pixel is an acronym for "picture element". It is the smallest distinguishable and resolvable area in a digital raster image. A pixel is also the discrete location of an individual photo-sensor in a digital camera. No attempt was made to base the digitisation protocol on the nature of the image, leading to rather big images. Kirsch illustrates this fact by showing a sixth-century

[1] Quote from: S. Aubenas, 'The photograph in print. Multiplication and stability of the image'. In: M. Frizot (ed), *A new history of photography*. Köln (Könemann Verlagsgesellschaft) (1998), p 225.

[2] Quote from: Ching-Chih and K. Kiernan (editors), *Report of the DELOS-NSF working group on digital imagery for significant cultural and historical materials. Prepared for the national science foundation (NSF) digital library initiative and the European Union under the fifth framework programme by the network of excellence for digital libraries (DELOS)* 2003. Online available at: <http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalimaging/DigitalImaging.pdf

[3] See: R. Kirsch, 'SEAC and the start of image processing at the National Bureau of Standards'. In: *Annals of the history of computing* IEEE, vol. 20 (1998), p7-13.

mosaic that contains about 80 x 46 carefully coloured and shaped tiles. Digitising this mosaic even with more (100 x 58) square pixels results in an inferior image. A much higher number of pixels are required in order to reveal the details of the original mosaic.

Graphics files can be considered as files that store any type of persistent graphics data (as opposed to text, spreadsheet, or numerical data, for example), and that are intended for eventual rendering and display. This definition can be found in the "Encyclopaedia of graphics file formats" published in 1994[4]. In the encyclopaedia nearly 100 different file formats can be found.

There are a number of reasons why there are so many different graphic file formats. The first reason is that there are a number of fundamental different types of graphical data. Each type requires its own file format. Three broad categories can be distinguished[5]:

?? Raster data: a group of sampled values, in either 2-dimensional or 3-dimensional space, that represents an image or that can be processed into an image[6].

?? Geometry data: mathematical description of space, in either 2-dimensional or 3-dimensional space, that represents the components of an image[7].

?? Latent image data: non-graphical data that can be transformed into useful images by some algorithmic process.

The digital image as discussed in this paper in all cases is a 2-dimensional raster image file or bitmap.

The second reason for the existence of a wide range of graphic file formats is that several proprietary formats were developed to prevent usage beyond the control of the original developer. The Kodak Photo CD format is an example of a format that was initially proprietary, but at a later stage, Kodak (more or less forced by the market that demanded more open formats) permitted developers to use to the format specifications. Some graphics file formats were directed towards the usage on specific dedicated hardware and thus have a specific format.

The third reason for the substantial number of graphic file formats is the wide range of design principles that were used by developers. The two main issues that influenced the development of specific graphics file formats were the speed to process the image and the memory required to represent the image. Longevity or durability is never mentioned as a specific design goal of a graphics file format.

The design of a graphics file format is based on the memory, speed and circuitry components of the hardware systems targeted to use the data. The amount of available memory can affect the speed of data access. Graphics applications are real memory hogs. Device independence decreases processing speed and increases memory requirements. Sometimes hardware-specific formats increase the speed and require less memory. An example of this is the fax-machine.

1.1. Digital raster images

A raster image file or bitmap file consists of a matrix of discrete pixel values created by the CCD of an image capture device. The CCD (charged coupled device) is a light sensitive sensor on a chip to convert the analogue signal into discrete digital codes. The sampling interval of the image capture device determines how many pixels are stored in the raster image file. These pixels-codes can be written in a number of ways. The raster image files can contain additional data such as image description information or a colour palette.

The memory capacity in bytes required to store the pixels of a digital raster image can be expressed by the formula: (PH X PV X PD) / 8. PH is the number of pixels in the horizontal dimension, PV is the number of pixels in the vertical dimension and PD ('pixel depth') is the

[4] J.D. Murray, J.D. and W. vanRyper, *Encyclopedia of graphics file formats*. Sebastopol, CA (O'Reilly & Associates), 1994.

[5] Ref. C.W. Browne and B.J. Sheperd, *Graphics file formats. Reference and guide*. Greenwich (Manning), 1995.

[6] Raster data is also called Bitmapped data.

[7] Vector graphics data is an example of Geometry data.

number of bits required to code the colour of an individual pixel. Depending on the film speed the sampling resolution of photographic film is the equivalent of 2000 – 5000 pixels per inch[8]. In order to render a raster file on a computer screen or to print it on paper a bitmap stored in computer memory has to be processed. A high-resolution bitmap, for example, has to be re-scaled for a full-screen view on a standard computer monitor. Another example is the dithering algorithm required to print a raster file on paper. Other image processing topics include image restoration and image enhancement. Data compression algorithms are applied in order to reduce the file size. As bitmap files tend to be very big compression is often applied. Smaller files require less storage capacity and less network bandwidth. The JPEG file compression method is used on a wide scale, mainly because Web browsers are able to de-compress JPEG-images automatically. JPEG-compressed images do not have the discrete raster structure anymore and the compression method results in a loss of image quality. As JPEG is both a compression algorithm and a file format this can lead to confusion. It is, for instance, possible to use the JPEG compression method for the creation of TIFF image files. The TIFF image file format is discussed further on in this paper. Concerning the durability of bitmap files, image-processing algorithms can be subdivided into two groups:

?? Image processing algorithms that do change values of the pixels in the sequence of bytes or bitstream of the stored bitmap. Examples are: image compression and modification of the pixel values by digital filters and contrast and brightness adaptations. Image compression often also abandons the rastered structure of the bitmap.

?? Image processing algorithms that do not change the pixels in the basic bitmap. Examples are image viewing and image printing software that manipulates the pixels for optimisation for a specific output. These operations are carried out in RAM memory of the computer and do not permanently manipulate the pixel values in the bitmap.

Image processing algorithms for a specific purpose that change the pixel values in the digital raster file can better not be used, because the characteristics of the file are changed. For a number of reasons the use of compression algorithms is advised against. There is a risk that the compression method becomes obsolete and in case the bitmap gets corrupted it is very difficult to repair the image. An uncompressed image can be repaired much easier. Lastly application of image compression often leads to loss of image quality.

1.2. Digital surrogates and digital born images

Libraries, archives and museums typically deal with two types of raster images: digital surrogates and digital born images. Digital surrogates are digital versions of analogue original images, such as historical photographs or photomechanical prints. Digital born images are created directly with a digital capture device and the result images do not have a close relation with analogue originals. Concerning the digital preservation of the two types of raster image files there is a lot of similarity. In both cases a suitable image file format and digital storage medium has to be applied and the documentation or metadata relevant for digital preservation has to take the character and usage of the image into consideration.

For digital surrogates the intended use of the digital image and the features of the analogue original influence the settings of the digital capture device and thus the characteristics of the raster image file. In a lot of situations digital surrogate images will be used for faithful on-screen viewing or the creation of a reproduction on paper. As in the future the output requirements can change, it is evident that the quality and the longevity of the digital objects are important. This makes the digitisation of visual sources rather a cyclic process than a linear process. The cyclic approach implies an active archiving policy in order to facilitate specific future utilization, such as scholarly research or online viewing. The concept "use-neutral" is used to express the requirements of high quality digital objects created in a conversion project. With this approach, an image is digitised once, at the highest level of

[8] See: F. Frey and J. Reilly, *Digital imaging for photographic collections. Foundations for technical standards.* Rochester (Image Permanence Institute, Rochester Intstitute of Technology), 1999. Online available at: <http://www.rit.edu/ipi> p21.

quality affordable, and studio standards such as colour matching and contrast levels are set so that the image can be used for multiple applications.
To a great extent the properties of the digital capture device, such as a digital camera determines the features of the digital born image created with it. Some standards used by digital capture devices should be taken into consideration for the digital preservation of digital born images, such as the EXIF standard for storing interchange information in image files, that is supported by almost all digital cameras. This issue is covered further on in this paper.
In general the creation of high quality digital raster images, or digital master images, requires a lot of resources. The loss of the digital images can imply a loss of valuable content. This justifies the required investments to guarantee long-term access to the digital images.

2. Preservation of digital objects

Until recently the vulnerability of digital objects was not considered as a problem. As long as digital objects are worthwhile to keep alive it is just a matter of moving it from to a new storage medium: from floppy disk to CD-ROM, from hard disk to tape. Also sometimes the data format must be converted in order to avoid that the data cannot be processed any more.
The Scientific American article “Ensuring the longevity of digital documents” by Jeff Rothenberg, published in 1995, is a widely cited publication that started to raise general awareness for the problem that digital documents have a rather short life[9]. Rothenberg wrote: that digital media “will last forever – or five years. Whichever comes first”. But it is not only the storage medium that raises concerns. The future understanding of the digital data is also of importance. What is the meaning of the bitstream on the storage medium and how can this meaning be interpreted in the future?

2.1. Digital preservation strategies

From 1995 onwards several digital preservation projects and studies were carried out on a wide range of subjects. They consisted of inventories and assessments of digital resources, tools and methods to preserve digital material and standards and guidelines to support digital preservation. Digital preservation refers to all the actions required to maintain access to digital materials beyond the limits of media failure or technological change. By the year 2000 three main strategies towards digital preservation have been developed[10]:

?? The technology preservation strategy. Preservation of the original software and hardware that was used to create and access the information. It involves preserving both the original operating system and hardware to run it.
?? The technology emulation strategy. Future computer systems emulate older, obsolete computer platforms as required. Emulation is the process of imitating obsolete systems on future generations of computers.
?? The digital information migration strategy. Digital information is re-encoded in new formats before the old format becomes obsolete. The purpose of migration is to preserve the intellectual content of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.

The existing consensus on the available strategies for digital preservation has not resulted yet in a common ground on how to implement these strategies and which preservation strategy allies with what type of digital material. Currently a number of experiments and feasibility studies are being carried out.
Another observation to be made is that the three digital preservation strategies are applied to a wide range of digital materials. To mention some: databases, computer programs, digital images, electronic texts, and web pages. The background and perception of the people implementing the digital preservation strategies determines how the digital materials actually

[9] J. Rothenberg, ‘Ensuring the longevity of digital documents’. In: *Scientific American*. (jan. 1995), p42-47.
[10] For more background information on Digital Preservation, see: M. Jones and N. Beagrie, *Preservation management of digital materials.* London (The British Library), 2001.

are understood and classified. Is a web site a form of an electronic text? Is a database inextricably connected with its database management system? Is it enough to preserve the result of a computer calculation or should the algorithms as such be preserved as well?
As a result of this differentiation of perception of digital materials a wide range of projects and research is carried out, sometimes with fundamentally different approaches while the character of the digital material is the same.
Digital preservation is a relatively young field of research and only future generations can judge whether digital preservation strategies implemented today were the right ones.

2.2. The Open Archival Information System (OAIS)[11]

Recently an International Standard was established that is gaining a lot of influence in the digital preservation research field. This standard, the “Reference Model for an Open Archival Information System” (OAIS) is developed under the direction of the “Consultative Committee for Space Data Systems” (CCSDS). The reference model is adopted as ISO 14721:2003 and establishes a common framework of terms and concepts relevant for the long-term archiving of digital data. An OAIS is defined as “an archive, consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community”. A Designated Community is defined as “an identified group of potential consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities”. The OAIS model is already widely used as a foundation stone for a wide range of digital preservation initiatives. The OAIS model can be considered as a conceptual framework informing the design of system architectures, but it does not ensure consistency or interoperability between implementations. The OAIS reference model contains three key high-level concepts:

1. The environment of an OAIS. An OAIS or archive is surrounded by “Producers” (which provide the information to be preserved), “Consumers” (which interact with OAIS services to find and acquire preserved information of interest), and “Management” (who set the overall OAIS policy as one component in a broader policy domain).
2. OAIS Information. Digital content is transported from Producer to Archive, and from Archive to Consumer in the form of “Information Packages”. Representation Information is required in order to understand the Data Object (this is either a physical object or a digital object) that is archived. Representation Information is the information that maps a Data Object into more meaningful concepts. Thus, an Information Object consists of two components: the Data Object and the Representation Information. An OAIS consists of a number of Information Packages, a conceptual container of two types of information: Content Information and Preservation Description Information (PDI). The PDI is divided into four types of preservation information called “Provenance”, “Context”, “Reference” and “Fixity”. It is necessary to distinguish between an Information Package that is preserved by an OAIS and the Information Packages that are submitted to, and disseminated from, an OAIS. These variants are referred to as the “Submission Information Package” (SIP), the “Archival Information Package” (AIP), and the “Dissemination Information Package” (DIP).
3. High-level external interactions. Producer and consumer interaction with the OAIS is based on specific Information Packages. A Producer delivers a “Submission Information Package” (SIP) to the OAIS for use in the construction of one or more AIPs. A Consumer receives a “Dissemination Information Package” (DIP), derived from one or more AIPs in response to a request to the OAIS.

[11] *Reference model for an Open Archival Information System (OAIS)*, published by Consultative Committee for Space Data Systems, CCSDS 650.0-B-1, Blue Book, January 2002. Available as ISO 14721:2003 standard. Online available at < http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>

The OAIS functional model consists of six archival functions:

1. Ingest. Contains the services and functions that accept the Submission Information Packages (SIP) from producers, prepares the Archival Information Packages (AIP) for storage, and ensures that the AIP and their supporting Descriptive Information become established within the OAIS.
2. Archival Storage. Contains the services and functions used for the storage and retrieval of the Archival Information Packages (AIP).
3. Data Management. Contains the services and functions for populating, maintaining, and accessing a wide variety of information.
4. Administration. Contains the services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis.
5. Preservation Planning. Contains services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete.
6. Access. Contains the services and functions, which make the archival information holdings and related services visible to Consumers.

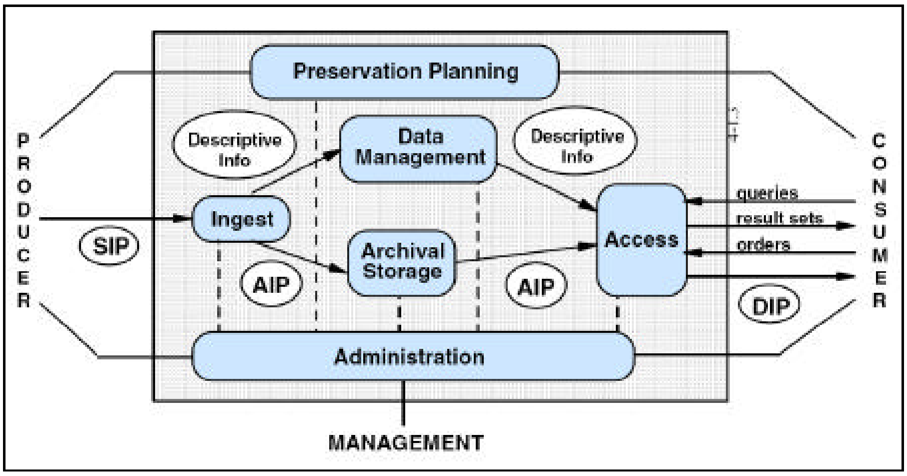

Figure 1: OAIS Functional Entities (source: OASIS / ISO14721:2003 p4-1).

Figure 1 contains both the three high-level concepts and the six entities of the OAIS reference model. Three areas of influence of the OAIS model can be distinguished:

?? First of all the model is used for the compilation of preservation metadata schema. Background information on Preservation Metadata is given further on in this paper

?? In the second place the OAIS model plays a role in the architecture and design of digital preservation information systems. The OAIS standards states: “It is assumed that implementers will use this reference model as a guide while developing a specific implementation to provide identified services and content”.

?? In the third place the OAIS model is used as a basis for conformance and many digital preservation information systems claim OAIS compliance. But a general accepted OAIS certification process does not exist yet. The Research Libraries Group (RLG) and the US National Archives and Records Administration (NARA) have established a task force on digital repository certification. Its purpose is to produce certification requirements for establishing and selecting reliable digital information repositories [12].

The OAIS standard distinguishes six mandatory responsibilities that an organisation must discharge in order to operate an OAIS archive. The OAIS must:

1. Negotiate for and accept appropriate information from Information Producers.
2. Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.
3. Determine which communities should become the Designated Community and, therefore, should be able to understand the information provided.
4. Ensure that the information to be preserved is independently understandable to the Designated Community.
5. Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.
6. Make the preserved information available to the Designated Community.

There is growing recognition that different kinds of digital data captured in different ways for long-term preservation will need various kinds of support. Highly structured digital materials tend to be inherently easier to preserve and access over time. Less structured materials tend to be harder to manage. Another way to categorize inherent persistence is whether the materials are homogeneous. This means closely tied to known and consistent rules regarding structure, technical parameters, and metadata.

2.3. Preservation metadata

Without documentation a digital object is just a sequence of binary digits. The longevity of digital objects will improve in case documentation on the digital objects is available. This is because information on issues such as the meaning of the bitstream that makes up the digital object, bibliographic information and data on the formal characteristics of the object will inform people and systems on the content, value and possible usage of the digital object. Thus, documentation on the digital object, or metadata, is an important facilitator for the longevity of the object.

Literally metadata means “data about data”. The library and information science literature published in the past years provides a significant role for metadata for the management of digital objects[13]. Metadata can be considered as documentation that gives information on the characteristics of ‘things’, both analogue and digital. Increased attention for the importance of metadata is caused by the rise of Internet. Internet is not a well-organised and structured library. The objects in the Internet information space can only be discovered in case metadata on the objects is available. In a library it is possible to search among collections. The metadata can have a wide range of functions. It can be used to identify versions of an object, to certify the authenticity, to indicate the status, to control the intellectual property rights, to mark the content structure, etc. For these and other functions a wide range of initiatives, projects, standards and guidelines are available[14].

Metadata is not only important for the discovery of resources, but also for its preservation. Information about the technological and other contexts of a digital object’s creation and use is known under the name ‘preservation metadata’. The OAIS reference model refers to metadata as “Representation Information”, information that maps an encoded digital object into more meaningful information. This encompasses the syntactic, structural, and semantic aspects of an encoding. Representation Information not only applies to digital objects, but it can also apply to the formats in which objects are encoded.

[12] The website of the RLG/NARA task force on Digital Repository Certification can be found at: <http://www.rlg.org/longterm/certification.html>

[13] See e.g.: M. Day, ‘Metadata for digital preservation: a review of recent development’. In: *Proceedings of the 5th European Conference for Digital Libraries* . Darmstadt (Springer), 2001, p161-172.

[14] For one specific function of metadata, resource discovery, a standard is used on a wide scale: the ‘Dublin Core Metadata Element Set’ (DCMES), see: <http://www.dublincore.org>. The fifteen core elements of DCMES are applied in a wide number of projects and initiatives in order to enable the discovery of objects on Internet.

There is no general accepted preservation metadata format that facilitates the longevity of digital raster images. A number of communities designed sets of data elements that can be used to improve the longevity of digital raster images. Some relevant sets of data elements are presented in the next section of this paper. Also the characteristics of the data elements used, such as the formulation of the definition and data type of the terms used differs very much.

2.3.1. METS: Metadata Encoding and Transmission Standard

An emerging method to store and express metadata, especially for audiovisual sources and their digital representations is the usage of a “wrapper” according to the METS specifications. METS, the Metadata Encoding and Transmission Standard[15], deals with the multiplication of metadata element sets in the recent years. In that sense it is a kind of meta-metadata. METS can be considered as integral part of the OAIS reference model (see section 2.2) and serves to facilitate the exchange of digital objects from one repository to another[16]. A METS document consists of seven major sections:

1. METS header: the header describes the METS document itself, e.g. information on the creator.
2. Descriptive metadata: may point to descriptive metadata external to the METS document, e.g. a record in a catalogue. Descriptive metadata may also be internally embedded.
3. Administrative metadata: provides information regarding how the digital objects were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of the files comprising the digital library object. As with descriptive metadata, administrative metadata may be either external to the METS document, or encoded internally.
4. File Section: lists all files containing content that comprise the electronic versions of the digital object.
5. Structural map: heart of a METS document. It outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element.
6. Structural links: allows METS creators to record the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map. This is of particular value in using METS to archive Websites.
7. Behaviour: can be used to associate behaviours of executables (computer programs) with content in the METS object. This section is not relevant for digital raster images.

METS uses the XML Schema language[17]. An XML Schema defines the allowable contents of an XML document. With METS a collection of related digital objects, e.g. digitised pages of a book or digitised photographs from an album, can be joined together. METS has a liberal approach towards the format of the data elements that describe the digital objects: any format can be used. Systems supporting the METS standard are still in the prototype phase. Figure 2 contains an example of a small part of a METS document, namely of the File section. The example makes clear that four digital images belong to each other. The four images could be digital raster images of a series.

[15] For more information on METS, see: <http://www.loc.gov/standards/mets>.

[16] Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) as explained in the OAIS Reference Model.

[17] See: <http://www.w3c.org/XML/Schema>

```
<METS:fileGrp>
<METS:file GROUPID="129131-1" MIMETYPE="image/tiff" ID="_79926" SEQ="1"></METS:file>
<METS:file GROUPID="129131-2" MIMETYPE="image/tiff" ID="_79927" SEQ="2"></METS:file>
<METS:file GROUPID="129131-3" MIMETYPE="image/tiff" ID="_79928" SEQ="3"></METS:file>
<METS:file GROUPID="129131-4" MIMETYPE="image/tiff" ID="_79929" SEQ="4"></METS:file>
</METS:fileGrp>
```

Figure 2: Small part of (of the “File” section) of a METS document, stating that 4 digital images (in TIFF format) belong to the same group.

2.4. Digital file format preservation

The way the binary digits are arranged in a digital file depends on the file format. Information on the internal syntax and semantics of the file format is important in order to understand and process the digital file. Format Registries that contain representation information about digital formats can help to ensure long-term access to digital files. A format registry can be used to identify, validate, characterise, transform and deliver digital objects, also in the long run.

The Global Digital Format Registry (GDFR) is an example of an initiative that investigates the possibilities to establish a sustainable format registry[18]. The provisional data model for the GDFR includes properties of the Registry itself and on properties of the format. The format properties are subdivided in descriptive properties, technical properties, system properties and administrative properties. The data model design was driven by consideration of the question: “What information would you want to have today to deal with a digital artefact from 50 years ago?” A proof of concept prototype of the GDFR is under development, but we are still far from an operational production registry.

The National Archives in the UK recently started a file format registry under the name PRONOM[19]. As stated on their website “PRONOM is an online source for information about file formats and software products. It is a resource for anyone requiring impartial and definitive technical information about the file formats used to store electronic records, and the software products that are required to create, render, or migrate these formats”. Currently the PRONOM system holds very limited information on data formats.

Next to Format Registries tools are developed to perform format-specific identification, validation, and characterisation of digital objects. Identification is the process of determining the specific format of a digital object. Validation is the process of determining the conformance of a digital object to the specifications for its purported format. Characterisation is the process of extracting preservation information (ref. OAIS-model) or metadata from an object. Whenever external metadata is submitted to a repository in association with digital objects it should be checked for consistency.

JHOVE (JSTOR/Harvard Object Validation Environment) is an extensible framework for this format-specific identification, validation, and characterisation of digital objects[20]. The JHOVE program currently available contains modules for a number of digital raster image formats, such as the common digital image formats GIF, JPEG, TIFF and PDF.

The Format Registry and digital object identification, validation and characterisation very well fits in the OAIS reference model (see section 2.2), mainly related to the Producer and Archive entities.

2.5. Longevity of storage media for digital objects

In 1995 the U.S. Department of Defence asked the National Media Laboratory to carry out a research on the life expectancy of storage media for digital data. The actual life expectancy of a particular storage medium depends upon the quality of the media manufactured, the number of times it is accessed over its lifetime, the care with which it is handled, the storage temperature and humidity, the cleanliness of the storage environment, and the quality of the

[18] More information on the GDFR and links to references can be found at: <http://hul.harvard.edu/gdfr>.

[19] The PRONOM System is accessible via <http://www.nationalarchives.gov.uk/pronom/>.

[20] The JHOVE software is made available publicly under the GNU General Public License (GPL) from the project website: <http://hul.harvard.edu/jhove>.

recorder used to write to the storage medium[21]. The research considered magnetic tape, optical disk, paper, and film media types. The two main factors influencing the life expectancy are storage temperature and relative humidity of the air. A storage temperature of 10 degrees Celsius and a relative humidity of 25% guarantee a reliable life expectancy of at least 20 years for both magnetic Digital Linear Type (DLT) and CD-ROM as optical disk. The best vendors of these products can deliver media that have a life expectancy of at least 100 years. Assumed is that new media is used, that the media is accessed infrequently, that the media is consistently stored at the indicated environmental conditions and that the storage environment is clean and free of dust, smoke, food, mold, direct sunlight, and gaseous contaminants.

Despite the fact that paper and microfilm in general have a longer life expectancy than optical disk and magnetic tape the durability of digital data expressed as collections of bits and bytes will be good enough for the reliable storage for a century. The ISO 18921:2002 standard is available to estimate the life expectancy of CD-ROM based on the effects of temperature and relative humidity[22]. The purpose of the standard is to establish a methodology for estimating the life expectancy of information stored on CD-ROMs. This methodology provides a technically and statistically sound procedure for obtaining and evaluating accelerated test data. An important measurement to determine whether a CD-ROM is still accessible is the “block error rate” or BLER. This is the ratio of erroneous blocks measured per second at the input at the data decoder.

It can be concluded that reliable media are available to store digital data for a long time. Hardware to access the bitstream on the media will probably become obsolete at an earlier stage. Monitoring of the available hardware to read the media is as important as monitoring the storage media. A bigger risk to loose the digital data is caused by the fact that the interpretation and processing of the data requires applications that can become obsolete. The durability of the data format is of higher importance than the durability of the storage medium. Recently the report “Care and handling of CDs and DVDs: A guide for librarians and archivists” was published. This report describes in a non-technical language the various types of CDs and DVDs in use, how they are made, and how they work. The report also contains current industry knowledge about media longevity, the conditions that affect life expectancy, and how to care for optical media[23].

3. Digital preservation of digital raster images

This section discusses issues relevant for the long-term access of a specific digital object, namely digital raster images and consists of two parts. The first part of this section concerns the available image file format standards and tries to answer the question which image file format standard is the most durable. The second part of this section elaborates on a number of important metadata element sets that support the durability of a digital raster image.

3.1. Standard image file formats

An obvious way to create durable digital objects is to use standardised data formats. A standard has the connotation of a well designed, widely used, and broad supported object. Empirically the requirements for standard data formats are given followed by an assessment of existing data formats for digital raster images. A standard data format for digital objects must meet three conditions:

1. A large community must use the data format during a considerable period of time. Making a data format obsolete that is used by a large community will have a negative influence on the reputation of the organisation that created the data format. The organisation will probably take the substantial user community into consideration when a data format has to be re-designed.

[21] See: C. Dollar, *Authentic electronic records: Strategies for long-term access*. Chicago (Cohasset Associates), 2000, p215.

[22] ISO 18921:2002 *Imaging materials – Compact discs (CD-ROM) – Method for estimating the life expectancy based on the effects of temperature and relative humidity*, International Organisation for Standardisation.

[23] See: F. Byers, *Care and handling of CDs and DVDs: A guide for librarians and archivists.* Washington (Council on library and information resources) 2004. Online available at: < http://www.clir.org/pubs/reports/pub121/pub121.pdf>

2. The specifications of the data format must be in the public domain or be published and supported by a standards developing organisation (SDO) such as ISO
3. A wide range of relevant systems has to support the format. E.g. a wide range of image capture devices as well as image processing systems must support a standard digital image data format. Cross-platform functionality of the data format is also a feature of this requirement.

For the specific type of digital objects, namely digital raster images, another three requirements for a standard data format can be formulated. These requirements are based on the principle that the data format must enable the creation of high quality digital raster images:

4. Data compression is not allowed for two reasons. Data compression will lead to loss of image quality and a compressed digital image has a bigger risk to become unreadable than an uncompressed digital image. Both issues are clarified.
   Raster images tend to be very big, so data compression algorithms are applied on a wide scale in order to decrease the data storage. Most data compression algorithms for raster images are based on the principle that the human eye cannot discriminate between all individual colours that are represented in an image. By giving closely related colours of the spectrum the same code, the number of required data codes can be reduced and thus the file size. In almost all cases compression of digital images of photographs will lead to loss of quality. Efficient data compression is a very important design issue for newly developed data formats for bitmap images. The recently developed JPEG2000 standard is an example of this[24].
   A corrupt bit in a compressed image file results in a “dead image”, as chances are that a corrupt bit in an uncompressed image is just a “dead pixel”. Thus, an uncompressed raster image is considered more durable than a compressed raster image, because the former probably can still be interpreted in case some bits are altered in due time.
5. A durable data format for digital raster images should contain facilities to store preservation metadata. The quality and granularity of the metadata is an important factor for the future usage of the data format and thus its longevity.
6. A durable digital image data format must enable the coding of specific significant characteristics, e.g. all colours, details and the dynamic range of an original or scene captured.

A considerable amount of graphic file formats is developed and also in the future new formats will be introduced. In order to determine which file format most closely meets the six requirements given above a number of graphic file format standards are evaluated and compared.
One of requirements of a durable digital raster file format is that the format must exist for a considerable period of time. The list of file formats stated in the *Encyclopedia of Graphics File Formats*[25], published about ten years ago, is used as a reference to raster image file format standards that are potential relevant. Four raster file formats mentioned in the book are still used today. These file formats are TIFF (tagged image file format), JPEG (joint photographers expert group), GIF (graphic file format) and PNG (portable network graphics).
Three file formats for raster images mentioned in the book are used on a rather wide scale, namely TIFF, JPEG and GIF. JPEG has a very big user community, because all standard Web browsers support it. Web browsers also support GIF, but this file format uses a patented compression algorithm and is less widely used on the web. The PNG format seems to be very appropriate to act as a file format standard for durable digital images, mainly because an independent group designed it. But PNG by default applies a (loss less) compression algorithm and the user community of PNG is not very big. To what extent the four file formats do meet the durability requirements is stated in Table 1[26].

[24] The website of the JPEG2000 standard can be found at: <http://www.jpeg.org/JPEG2000.html>.

[25] See note 4. Both the first and second edition of the Encyclopedia is used. The first edition is published in 1994, the second in 1996. Until now a update is not published.

[26] Graphics file formats that are not available on several platforms, such as Microsoft Windows Bitmap (BMP) and graphics file formats that are part of an imaging system, such as the Kodak Photo CD (PCD), are excluded from the selection.

	Raster file requirements	TIFF	JPEG	GIF	PNG
1	Used by a large community over a long time	+	+	+	-
2	File format specification is published	+	+	+	+
3	Supported by a wide range of applications	+	+	+	+
4	Supports un-compressed / single page images	+	-	-	-
5	Facilities for preservation Metadata	+	-	-	+
6	Enable “full informational capture”	+	-	-	+

Table 1: Durability specifications and digital raster file formats.

The TIFF image file format seems to be the most durable standard to be used for the coding of digital images that are high quality digital surrogates of historical photographs.

3.1.1. The TIFF image file format

The raster image file format TIFF (‘tagged image file format’) meets the first three requirements for durable digital objects mentioned in above. The file format is largely hardware and software independent and has been around for more than 10 years. The format is used for the storage of raster images by all digital conversion initiatives in the cultural heritage sector that have the ambition to create high quality digital master files. The specification of the most recent version of the TIFF data format is freely available via the Website of Adobe[27]. TIFF version 6 was made public available in 1992. The original TIFF specification was released in 1986 by Aldus Corporation that later was acquired by Adobe, as a standard method of storing black-and-white images created by scanners and desktop publishing standards. The functionality of the subsequent versions of the TIFF raster image file format improved enormously. TIFF version 6.0 supports the coding of colour, compression methods and metadata. There is no indication that a new version of the file format will be released. If it were, it can be expected that this new format will be backward compatible with earlier versions. Rumours on graphics newsgroups archives on Internet state that Adobe is not happy with the inheritance of the TIFF standard when it acquired Aldus, because it is a competitor of the PDF standard developed by Adobe. Adobe cannot abandon the support of the TIFF standard because its user community is very big.

The success of TIFF as a widely used standard for digital raster images is caused by its extensible nature and support for numerous data compression schemes. So developers are able to customise the format to fit any peculiar data storage needs. Both compressed and uncompressed images can be coded in the TIFF standard and it is also possible that more than one image is stored in a file formatted according to the TIFF standard. The TIFF standard allows the inclusion of an unlimited amount of private or special-purpose information, e.g. preservation metadata. This means that features that restrain the longevity of digital images as stated above, are also part of this standardised file format.

A drawback of the TIFF standard is that Web browsers do not support it. A helper application, plug-in or conversion to another file format (e.g. JPEG) is needed before a Web browser can process an image. A TIFF file cannot have more than 4 gigabytes of raster data.

Baseline TIFF

The TIFF version 6.0 specification is divided into two parts: baseline TIFF and TIFF extensions. Baseline TIFF is the core of TIFF, the essentials that all mainstream TIFF developers should support in their products. TIFF extensions are TIFF features that may not be supported by all TIFF readers and this can hamper successful interchange and thus lowering the durability of the digital image.

A TIFF file starts with an 8-byte image file header. The first two bytes of this header define the byte order used within the file. The second and third bytes contain an arbitrary number in carefully chosen format, also called magic number, that identifies the file as a TIFF file. The last four bytes of the file header contain a reference to the location of the first Image File

[27] The specifications of the TIFF file format version 6.0 can be found at: <http://partners.adobe.com/asn/tech/tiff/>

Directory (IFD). This “byte offset” always refers to a location with respect to the beginning of the TIFF file. The IFD is the second section of a TIFF file and contains information fields or tags that are described below. The third section of a TIFF file contains the bitmap data.

An essential feature of the TIFF standard is that it consists of fields that contain information on the bitmap data. This information is required by image processing systems in order to render the image. Other fields are used to store textual documentation on the image. Baseline TIFF has 36 fields. These are given in table 2 in ascending order by decimal code.

For digital raster images two types of TIFF images are relevant: greyscale and full colour images. Whether the greyscale or full colour image is applied is determined by the digitisation requirements. Baseline TIFF sets a number of required fields for these image types that can be found in the third column of table 2.

A baseline TIFF file containing a greyscale image requires 11 information fields. A baseline TIFF file containing a full colour image requires 12 information fields. The information field “SamplesPerPixel” is optional for greyscale images, but is required for full colour images. The actual location of the data in a TIFF file is rather complex and three required information fields in a baseline TIFF manage the location of the image data. These information fields are: “StripOffsets”, “RowsPerStrip” and “StripByteCounts”.

Baseline TIFF supports a small number of data compression methods, coded in information field 259. As durable images should not be compressed this information field must have the value “1”, meaning “No compression”. The remaining required information fields of the baseline TIFF specification serve the characteristics of the pixels that make up the raster data. A fixed value for photometric interpretation is not required. The value determines whether “0” is imaged as white (value = 0) or “0” is imaged as black (value = 1).

The fifth column of table 2 indicates whether the data in the information field can be considered as preservation metadata for digital raster images, either greyscale or full colour. Preservation metadata is documentation that helps future users of the image (people and systems) to understand and process the image. The more documentation the better, but information fields that are not relevant for the image type or information fields that hamper the digital durability, e.g. information fields that support multi page documents, as well as information fields that are classified as “not recommended for general interchange” by the TIFF 6.0 standard are excluded from this list.

Table 2 gives 24 information fields of baseline TIFF that support digital durability. The application of five information fields hampers digital durability and seven information fields are not relevant for digital durability.

Tagname	*Decimal code*	*Required for grayscale / full color images*	*Value (if applicable for greyscale and colour images)*	*Preservation metadata*	*Usage hampers durability*
NewSubfileType	254				X
SubfileType	255				X
Imagewith	256	X	Number of pixels	X	
Imagelength	257	X	Number of pixels	X	
BitsPerSample	258	X	8 for greyscale and 8 8 8 for colour	X	
Compression	259	X	1 = “uncompressed”	X	
Photometric Interpretation	262	X	0 or 1 for greyscale / 2 for full colour	X	
Thresholding	263				
CellWidth	264				
CellLength	265				
FillOrder	266				
Image Description	270			X	
Make	271		Scanner manufacturer	X	
Model	272		Scanner model	X	
StripOffsets	273	X		X	
Orientation	274		Baseline TIFF only supports value “1”		
SamplesPer Pixel (required for full colour images)	277	X	“1” for greyscale images and (optional) “3” for full colour images (RGB)	X	
RowsPerStrip	278	X		X	
StripByteCounts	279	X		X	
MinSampleValue	280		Contains min. / max. pixel value for statistical purposes	X	
MaxSampleValue	281			X	
XResolution	282	X	Number of pixels per resolution unit (=tag 296)	X	
YResolution	283	X	Number of pixels per resolution unit (=tag 296)	X	
Planar Configuration	284		Baseline TIFF only supports value “1”		
FreeOffsets	288				X
FreeByteCounts	289				X
GrayResponse Unit	290			X	
GrayResponse Curve	291			X	
ResolutionUnit	296	X	1 (=none) or 2 (= inch) or 3 (=cm.)	X	
Software	305			X	
DateTime	306		date / time of image creation	X	
Colormap	320		Only relevant for palette colour images		
Artist	315			X	
HostComputer	316			X	
ExtraSamples	338				X
Copyright	33432		Copyright notice	X	

Table 2: Assessment of Baseline TIFF 6.0 information fields.

In principle the baseline TIFF version 6.0 file format meets all six requirements for durable raster images. The TIFF 6.0 extensions are analysed below in order to determine whether better methods are available for the formulation of preservation metadata and the coding of all required significant characteristics.

TIFF extensions

TIFF extensions are TIFF features that may not be supported by all TIFF readers. The official TIFF 6.0 standard contains a number of TIFF extensions and there are a number of TIFF extensions that are published independently from the officially published TIFF 6.0 standard.

The officially published TIFF 6.0 specification contains four groups of extensions. One group of extensions concern data compression methods. Another group covers an alternative for the organisation of the image in tiles instead of strips. The third type of extension improves the quality of a specific image type, namely halftone images. The only extension relevant for the durability of a digital image is the TIFF 6.0 extension for better colour management. The CIELAB colour space, supported by the extension on the TIFF version 6.0 specification, has excellent applicability for device-independent manipulation of continuous tone images. For digital surrogates of colour photographs e.g. this high quality colour space is of great importance. In case the CIELAB colour space is used the information tag 262 (Photometric Interpretation) should have the value "8". This value has the meaning "1976 CIE L*a*b".
Next to extensions that are part of the official TIFF standard there are also a number of separate published extensions. To mention some examples: a specification of TIFF especially for GIS applications, a TIFF specification that support the JPEG compression method and a TIFF extension as pre-press interchange format.
In principle the TIFF specification meets the criteria to act as the standard for the creation of durable digital images, but features that obstruct durability, such as compression and inclusion of non-standard private tags, should not be used. This means that the TIFF standard is rather sloppy and tolerant where it comes to compliance to the standard[28].
There is one extension on the TIFF 6.0 format that has the status of an international ISO standard that can be considered as a durable file format: the TIFF/EP (TIFF/Electronic Photography) image data format TIFF/EP is developed as a standard for the coding of images of electronic still-picture cameras and is based on TIFF version 6.0, but TIFF/EP contains a number of new tags[29]. TIFF/EP is also known under the name EXIF (Exchangeable image file format). This format plays an important role in the formulation of metadata and is discussed further on in this paper.
The TIFF 6.0 specification can be considered as durable, provided that it is applied in a specific way. No compression method should be applied and no multiple page images must be created. TIFF 6.0 contains a number of information tags that enable the storage of preservation metadata. But support for storage of preservation metadata in baseline TIFF is rather limited. Only a limited amount of information fields is available for the storage of preservation metadata. Next to that the scope and purpose of these information fields is described rather vague. E.g. the requirements of the information field "ImageDescription" are not specified. The coding of all significant characteristics is also problematic, because baseline TIFF does not support accurate coding of colour information. The TIFF 6.0 extension contains a high quality colour space (CIE LAB) that supports the accurate coding of colours.

3.2. Preservation metadata relevant for the longevity of digital raster images

Preservation metadata is documentation that plays a role in the long-term access of digital objects. The OAIS reference model (see section 2.2) can be used for the formulation of the data elements that users in the future, the Designated Community, need in order to understand and process the digital object. Currently a number of sets of data elements, or schemas, do exist that can play a role in the longevity of digital raster images. An example is the SEPIADES schema aimed to catalogue a wide number of characteristics of photographic collections[30]. Also standards for the representation and communication of bibliographic information, such as MARC[31], can be used for the formulation of preservation metadata of "non-book" material. The goal of the PREMIS project is to develop best practices and recommendations for implementing preservation metadata for digital objects[32]. Results from the PREMIS project can be expected later in 2004.

[28] The PNG raster image file format (Portable Network Graphics), a W3C recommendation is considered as a well-designed, easy accessible, superior format, but after the promising outset, this format did not receive a wide user community. See: <http://www.w3c.org/Graphics/PNG>.
[29] See: ISO 12234-2:2001 *Electronic still-picture imaging – removable memory – Part 2: TIFF/EP image data format,* International Organisation for Standardisation.
[30] For more information on SEPIADES, see: E. Klijn (ed.), *SEPIADES. Recommendations for cataloguing photographic collections*. Amsterdam (European Commission on Preservation and Access), 2003.
[31] See: <http://www.loc.gov/marc>.
[32] PREMIS stands for: "PREservation Metadata: Implementation Strategies", see: <http://www.oclc.org/research/projects/pmwg/>. The project is based on the OAIS Reference Model.

In general data elements that are part of a metadata schema can be stored in three ways. In the first place data elements can be stored in the header of the image file, such as the information fields in a TIFF file (see section 3.1.1). Data elements can also be stored in the file system through the names of the directories and image files. In the third place data elements can be stored in a separate database.

A promising standardised construct to express preservation metadata is the METS "wrapper" (see section 2.3.1 of this paper). A METS document is an XML formatted document that contains all relevant metadata of a digital object or a set of related digital objects as well as its related analogue originals.

In the next section of this paper one type of preservation metadata, namely technical metadata for digital raster images is covered more in detail. Technical metadata is only a subset of the complete suite of preservation required for long-term access, but it has often been called the first line of defence against losing access. Technical documentation is relevant in two closely related fields. Firstly, technical metadata facilitates the smooth exchange of digital images between different systems. Secondly, a future migration process by copying images to new formats benefits from standardised technical metadata.

3.2.1. Technical Metadata for digital raster images

The information fields of the TIFF standard contain a number of data elements that can be considered as relevant for long-term access (see table 2). The set of TIFF information fields is extended by other standards. In the situation a non-TIFF image file format is used an alternative construct for the expression of preservation metadata is required. Two standards that are related to the TIFF standard are important for the expression of technical metadata of digital raster images: the EXIF standard and the NISO Z39.87 draft standard.

The EXIF standard and related DCF specification originates from the digital camera manufacturers community[33]. This standard is relevant for digital born images. The draft standard NISO Z39.87 "Technical Metadata for Digital Still Images" is initiated by the cultural heritage community and mainly intended for the formulation of technical metadata of digital surrogates of analogue originals[34].

EXIF an DCF

EXIF stands for "Exchangeable Image File Format", and is a standard for storing interchange information in image files, especially those using JPEG compression. The specification DCF (Design Rule for Camera File system) was drawn up for the purpose of simplifying the interchange of image files and related files on digital still camera and other equipment. DCF formulates the names of files and arrangement of directories. EXIF stores metadata in the beginning of the files und uses the standard colour space sRGB[35]. Most digital cameras now use the EXIF format and DCF specification, e.g. digital cameras manufactured by Canon, Kodak, Sony, and Olympus. The format is developed by the "Japanese Electronics and Information Technology Industries Association" (JEITA). The EXIF image file format was established with the aim of realising a common format for the image files used with digital still cameras and other related equipment, making these products more convenient for end users. With the rapidly growing popularity of digital still cameras, there are increasing demands for image file inter-changeability, which will allow images captured on one camera to be viewed on another, or output directly to a printer.

EXIF version 2.2, established in April 2002 specifies the structure of image data files and the information fields or tags used by the standard. The EXIF standard extends the mandatory TIFF information fields with additional EXIF tags. Figure 3 contains the technical metadata according to the EXIF standard as created by a common consumer digital camera. The metadata is showed as an XML file. The EXIF standard contains much more tags, but the camera uses not all of them. According to the EXIF standard the code "1" for the tag <ColorSpace> refers to the sRGB colour space.

[33] Details on EXIF and DCF can be found at: <http://www.exif.org>.

[34] NISO Z39.87-2002 "Data Dictionary – Technical Metadata for Digital Still Images" can be downloaded from: <http://www.niso.org/standards/resources/Z39_87_trial_use.pdf>

[35] More information on the sRGB colour space can be found at: <http://www.w3.org/Graphics/Color/sRGB.html>.

The EXIF standard specifies the following three date tags that all have a specific meaning:

?? <DateTime> records the date and time of file updating, like a file time stamp.
?? <DateTimeOriginal> records the date and time when an image was shot.
?? <DateTimeDigitised> has the date and time when digital data was created.

For a digital still camera, in many cases the contents for the three date tags are identical as can be seen in figure 3.

```
<Exif>
    <CameraManufacturer>Canon</CameraManufacturer>
    <CameraModel>Canon PowerShot A70</CameraModel>
    <Orientation>top, left</Orientation>
    <XResolution>1/180</XResolution>
    <YResolution>1/180</YResolution>
    <ResolutionUnit>Inches</ResolutionUnit>
    <DateTime>2004:07:21 12:51:34</DateTime>
    <YCBCrPositioning>Centered</YCBCrPositioning>
    <ExposureTime>1/60 sec</ExposureTime>
    <FNumber>4.0</FNumber>
    <ExifVersion>0220</ExifVersion>
    <DateTimeOriginal>2004:07:21 12:51:34</DateTimeOriginal>
    <DateTimeDigitized>2004:07:21 12:51:34</DateTimeDigitized>
    <BitsperSample>2</BitsperSample>
    <ExposureBiasValue>0.0</ExposureBiasValue>
    <MaxApertureValue>4.0</MaxApertureValue>
    <MeteringMode>Multi Segment</MeteringMode>
    <Flash>Unknown</Flash>
    <FocalLength>11.10 mm</FocalLength>
    <FlashPixVersion>0100</FlashPixVersion>
    <ColorSpace>1</ColorSpace>
    <Width>1536 pixels</Width>
    <Height>2048 pixels</Height>
    <SensingMethod>One-chip color area sensor</SensingMethod>
 </Exif>
```

Figure 3: EXIF Metadata in XML format (note: EXIF does contain much more information fields than stated in the figure).

Compressed image files are recorded as JPEG[36] with application marker segments inserted. Uncompressed files are recorded in TIFF version 6.0 format. Related attribute information for both compressed and uncompressed files is stored in the tag information format defined according to TIFF version 6.0. Information specific to the camera system and not defined in TIFF is stored in private tags registered for EXIF. The fact the EXIF supports the TIFF uncompressed image file format, does not mean that all EXIF compliant digital capture devices are able to create uncompressed digital images. A lot of consumer digital cameras are only able to process compressed image files.

DCF is aimed at the creation of a user environment in which consumers of digital images can combine products more freely and exchange media readily. DCF specifies rules for recording, reading and handling image files and other related files used on digital still cameras or other equipment. DCF is applicable to products for writing image files on an interchangeable storage medium. According to DCF the name of files and directories may only contain digits, the 26 characters from the Latin alphabet (no distinction between lower and upper case), or the character "_" (lower dash).

[36] The JPEG compression method is defined as standard: ISO/IEC 10918-1. See: <http://www.jpeg.org/jpeg/index.html>.

DCF consists of three specifications:

?? Media specification. (Specifies state of data on a storage medium)
?? Writer specification. (Specifies the recording function, e.g. by a digital camera)
?? Reader specification. (Specifies the playback function)

The DCF media standard defines the structure of the directory and the directory names on devices that store digital images. The directory with the name “DCIM” (Digital Camera IMages) directly under the root directory is called the DCF image root directory. The directories that store DCF objects are called DCF directories.

NISO Z39.87 Data Dictionary – Technical Metadata for Digital Still Images

```
<?xml version="1.0" encoding="UTF-8" ?>
<METS_Profile ...>
...
<extension_schema>
        <name>The MIX Technical Metadata for Still Images XML Schema</name>
        <URI>http://www.loc.gov/standards/mix/mix.xsd</URI>
</extension_schema>
...
<mix:mix>
        <mix:BasicImageParameters>
                <mix:Format>
                        <mix:MIMEType>image/tiff</mix:MIMEType>
                        <mix:ByteOrder>big-endian</mix:ByteOrder>
                        <mix:Compression>
                                <mix:CompressionScheme>1</mix:CompressionScheme>
                        </mix:Compression>
                        <mix:PhotometricInterpretation>
                                <mix:ColorSpace>2</mix:ColorSpace>
                        </mix:PhotometricInterpretation>
                </mix:Format>
        <mix:BasicImageParameters>
        ...
        <mix:ImageCreation>
        ...
        <mix:ImagingPerformanceAssessment >
        ...
</mix:mix>
...
</METS_Profile>
```

Figure 4: Data elements of the NISO Z39.87 standard expressed as the XML Schema “MIX” and included in a METS wrapper.

The purpose of the NISO Z39.87 data dictionary is to define a standard of data elements of digital images. The data dictionary has been designed to facilitate interoperability between systems, services and software as well as to support the long-term management of and continuing access to digital image collections. The intended audience of the NISO Z39.87 standards are cultural institutions, publishers, right holders, and other organisations engaged in digitising visual materials from archival collections. The data elements are structured to accommodate practices associated with digital copy photography, such as the use of technical targets, as well as the techniques to direct digital photography of original scenes.

The NISO Z39.87 data dictionary covers four categories of functions:

?? Basic image parameters record information crucial to displaying a viewable image.

?? Image creation data elements record information important for understanding the technical environment in which a digital image file was captured.

?? Imaging performance assessment data elements record information that allows evaluation of the quality of the digital image, or output accuracy.

?? Change history data elements record information about the process applied to an image over its life cycle.

The data elements of the NISO Z39.87 data dictionary build and expand on technical metadata available in other standards, such as the TIFF image file format version 6.0. An XML Schema that contains the data elements of the NISO Z39.87 standard is available. This XML Schema, “NISO Metadata for Images in XML Schema” (MIX), provides a format for interchange and/or storage of technical metadata[37]. Figure 4 contains a small part of a METS document that refers to the MIX XML Schema. The item “Format”, one of the items of the section “Basic image parameters”, consists of a number of data elements. According to the NISO Z39.87 data dictionary the value “1” for <CompressionScheme> stands for “Uncompressed” and the value “2” for <ColorSpace> means “RGB”.

According to the terminology of the OAIS reference model, technical metadata is part of Representation Information. In a preservation repository this information will become part of an OAIS Information Package. The METS specification consists of a number of objects that can be seen as an implementation of the OAIS Information Packages SIP, AIP and DIP (See figure 1). METS is an XML Schema that can be used for the encoding of descriptive, administrative, and structural metadata regarding objects within a digital library. The NISO Z39.87 data dictionary can be considered as a part of the administrative metadata of the METS specification. As an XML Schema implementation of the NISO Z39.87 data dictionary does exist under the name “MIX” (see figure 4) the three standards (OAIS, METS and NISO Z39.87), each relevant on a specific level for digital preservation of digital raster images, can be combined. This combination is illustrated in figure 5. An XML Schema of the EXIF specification would also fit within the Administrative Metadata section of a METS standard.

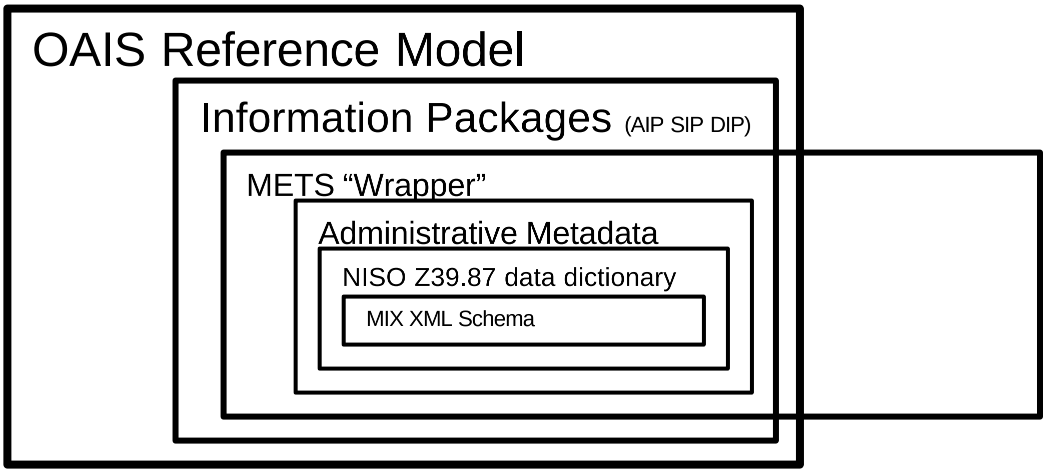

Figure 5: Relation between OAIS Reference Model, METS Schema and the NISO Z39.87 Data Dictionary.

[37] This XML Schema is referred to as “NISO Metadata for Images in XML”, and abbreviated as MIX. See: <http://www.loc.gov/standards/mix>.

Conclusion

Digital preservation is a relatively young field of interest. A number of principles and standards do exist upon which solutions can be based to enable the durability of digital raster images. The two most important issues that determine the long-term access to digital raster image are the application of a standardised image file format and the creation of preservation metadata. Based on an analysis of the requirements for a suitable standard image file format and a review of the image file formats that are used for about the last 10 years, currently the baseline TIFF version 6.0 seems to be the most durable raster image file format.
Concerning preservation metadata a number of categories can be distinguished. For the formulation and expression of technical metadata a number of solutions are available. The EXIF standard applied by a wide number of digital still cameras contains a number of data elements that are relevant for future access to digital images. The cultural heritage institutes aiming at long-term access to digital images specifically developed the NISO Z39.87 “data dictionary – technical metadata for digital still images”. The data elements of this NISO draft standard can be stored in the durable XML data format by using the MIX XML Schema that might be part of a METS document.
The contours of a digital preservation solution for digital raster images are available at the moment. The OAIS reference model is important as conceptual framework for the design of a digital archive. Proactive acting and commitment to keeping digital raster image accessible in the long run is also important. This paper contains relevant starting points for the implementation of a sound strategy to minimize the risk that digital raster images are not accessible anymore in the near future.