Permanent pixels

Building blocks for the longevity of digital surrogates of historical photographs

This thesis has been made possible with financial support from:

- Delft University of Technology (DITSE BETADE)
- Netherlands Institute for Scientific Information Services (NIWI-KNAW)
- Data Archiving and Networked Services (DANS)

Printing: GrafiSchoth
Cover Design: Erwin van Krey
Copy-editing: Lorna Cullen

Permanent pixels

Building blocks for the longevity of digital surrogates of historical photographs

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 1 november 2005 om 15:30 uur
door

Marinus Petrus Maria VAN HORIK

doctorandus in de geschiedenis
geboren te Aarle-Rixtel

Dit proefschrift is goedgekeurd door de promotoren:
Prof. dr. H.G. Sol
Prof. dr. H. Koppelaar
Toegevoegd promotor: Dr. K. van der Meer

Samenstelling promotiecommissie:

Rector Magnificus	Voorzitter
Prof. dr. H.G. Sol	Technische Universiteit Delft, promotor
Prof. dr. H. Koppelaar	Technische Universiteit Delft, promotor
Dr. K. van der Meer	Technische Universiteit Delft, toegevoegd promotor
Prof. dr. F. C. J. Ketelaar	Universiteit van Amsterdam
Prof. Dr. Ir. A. Verbraeck	University of Maryland, United States of America
Dr. P.K. Doorn	Data Archiving and Networked Services
Dr. S. Ross	University of Glasgow, United Kingdom

Preface

Science is not about finding the truth; it is about finding hypotheses that can be demonstrated to be consistent with reality. ([TAN98] p. 310)

The wish to write a dissertation has been latent ever since I started to work in the field of the application of information technology in the Humanities. Since early 1989, I have been involved in several activities in which information technology was applied to create, analyse and archive digital resources relevant for scholarly research in the Humanities. My involvement in the Netherlands Historical Data Archive (NHDA) – ironically often struggling for survival – made me realise how valuable and vulnerable digital research datasets are. My activities in a wide range of digital conversion projects in which a diversity of historical source material was digitised strengthened my feeling that the digital surrogates, often created at a high cost, run a great risk of being lost in the future. Increasingly, I have been intrigued by the problems related to long-term access to digital data, but I realised that digital preservation is a very wide field of interest.

The BETADE research programme of Delft University of Technology had a role to play in making my latent plans to start a PhD dissertation more concrete, as document management and longevity of digital documents was one of its application fields. My research proposal was focused on the longevity of a specific class of digital document, namely digital surrogates of historical photographs. My experiences concerning the digitisation of historical analogue sources, from 1997 onwards at Netherlands Institute for Scientific Information Services (NIWI-KNAW), taught me that the digitisation of continuous tone sources, such as photographs, is much more complicated and expensive than that of textual sources. By concentrating my research on the longevity of this specific type of document I expect to contribute to the realisation of long-term access to this type of digital object rather than to come to more general conclusions concerning digital preservation issues. Of course, extrapolation of the research results to more general digital preservation principles can also be expected.

Within the context of the problem of long-term access to digital objects in general this dissertation takes the longevity of digital surrogates of historical photographs into consideration. The focus on this specific type of digital data object has two foundations. First, there is the ubiquitous view among archives, libraries and museums that analogue originals and their digital counterparts are closely related. The features of a digital surrogate, such as a scanned historical photograph, are based on an assessment of the analogue original as well as the intended function of the digital surrogate, also in the long term. This gives the digital objects to be preserved a very specific character that justifies the scope of this research on the longevity of a specific digital data object.

The second reason for concentrating in this dissertation on a specific type of digital data object is more pragmatic. The focus on the longevity of digital surrogates of historical photographs enables the investigation of available building blocks for digital preservation such as guidelines, procedures, tools, standards, strategies and methods to enable long-term access to digital data objects. This dissertation examines in detail the role of digital preservation strategies, file format standards, content format standards, metadata schemas and other building blocks for the realisation of durable digital surrogates of historical photographs.

It is obvious that the longevity of digital objects is very much determined by the level of commitment of an organisation that considers itself as the curator of the digital objects. This is much less a technical than an organisational issue. Despite the fact that the organisational structure of the NHDA and its successor NIWI-KNAW turned out to be not very durable, the continual organisational commitment to provide access to digital objects relevant for research in the Humanities has prevented the loss of data archives up to now. Based on new insights, the recently founded organisation DANS (Data Archiving and Networked Services) took over and augmented the digital data archiving responsibilities of NHDA and NIWI-KNAW. It is important that this organisation, just like its predecessors, puts efforts into research into digital preservation in order to provide long-term access to its digital assets.

I have appreciated the stimulating supervisory instructions of Professor Henk Koppelaar and the comments of Professor Henk Sol. Kees van der Meer was my companion at Delft and helped me to acclimatise. His day-to-day advice, suggestions and discussions helped me on my way. I am very grateful to Peter Doorn of NIWI-KNAW and DANS, who from the outset motivated me to embark on the PhD research and who was very cooperative and flexible in establishing an agreement between the BETADE research programme of Delft University of Technology and NIWI-KNAW. I also very much appreciate the stimulating active involvement of Peter in my research. Furthermore I owe a great deal of gratitude to the other members of my promotion commission for their investment in my dissertation.

I would like to thank Charles Dollar, one of the founding fathers of digital preservation solutions of electronic records, who participated in a BETADE workshop and provided me with important initial research suggestions. I would also like to thank Alexander Verbraeck and Ajantha Dahanayake, supervisors of the BETADE research programme, and Frans Dondorp, my ‘roommate’ in Delft, for our thought-provoking discussions. For the indispensable practical experience concerning the digitisation and long-term archiving of digital photographic collections, I appreciate very much the collaboration in several projects of Edwin Klijn, Rutger Kramer, Yola de Lusenet, Laurents Sesink, Paula Witkamp, Douwe Zeldenrust and Joris van Zundert, all working at different research institutes of the Royal Netherlands Academy of Arts and Sciences.

And, finally, I owe much gratitude to Hilde and our daughters Hannah and Vera for their patience and support during the period of writing 'my book'.

Leiden, May 2005
René van Horik

Contents

1 Introduction

The problem of creating stable and permanent images was not solved with the invention of photography in 1839. This is illustrated by the following quote:

> ... 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 ([AUB98] p. 225).

In the closing decades 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:

> 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 ([CHI03] p. 4).

Obsolescence of the digital image file format and deterioration of the digital data storage medium are among the main factors that threaten long-term access to digital images. A number of digital preservation approaches came into existence to address the problem of digital objects not being accessible in the future. A digital image acting as a surrogate of an historical photograph is a specific type of digital object and the creation and durability of this digital surrogate is the main subject of this study.

This introductory chapter consists of three parts. In the first section a global overview is given of the digitisation activities in archives, libraries and museums from the time when the computer was introduced. This overview serves as the context for the application of information technology in institutes that preserve and disseminate cultural heritage. The second section of this chapter discusses the main common strategies to prevent digital objects becoming unusable in the future. In this section the longevity of a specific type of digital object, namely digital surrogates of historical photographs, is also introduced. In the third section of this chapter the research approach is clarified. This section contains the formulation of the research question, a discussion of the research philosophy, the research strategy and research instruments.

1.1 Digitisation in memory institutes

In Reference [MEE02] Van der Meer elaborates on the design, functions and characteristics of document management systems and provides insights into the way analogue source material held by organisations such as archives, museums and libraries can be converted into an electronic environment. For archives, libraries and museums the digitisation of analogue source material is an exciting way to open up and exploit their holdings. Digitisation is applied on a wide scale and on the Internet an enormous amount of archive, library and museum collections can be found. The UNESCO portal website refers to more than 16,000 digital archive and library collections.[1] The digital surrogates of analogue sources can be used for a wide range of purposes, such as a truthful representation of the original or a global reference to the original. Archives, libraries and museums may be considered as the 'memory of society', as their main role is to collect, select, store and provide access to all kinds of artefacts created by society. Obviously this memory must be accessible now and in the distant future, even in digital form.

The application of computer systems in memory institutes started in the 1970s with the creation of electronic catalogues. The analogue card catalogues that provided access to the collection items were replaced by databases. The bibliographic information on the catalogue cards was converted into database fields. The electronic catalogue provided automatic access to analogue objects such as books, works of art and documents. In the 1980s the conversion of printed source material, such as books and articles, into digital files started to become widespread. Two types of digital files were created. The first file type represents the storage medium of the original analogue document; this is often called a digital image. The second file type represents the coded content (for instance, characters and figures) of the original document and, depending on the type of original, can be an electronic text, table or database.

Initially, digital images of printed source material contained only two tonal values, representing the black ink and white paper of the originals. Compared with the 'full colour' images that were created at a later stage, these binary image files had a relatively small file size and did not require extensive computer power to be rendered. Then, in the 1990s the digitisation of visual sources with continuous tone scales came to the fore. These sources contain a wide range of intermediate shades of colour tones. Resources and technology became available for memory institutes to create high-resolution, high-dynamic digital images. The images very

[1] The website of the Archives portal and Libraries portal of UNESCO can be found at: <http://www.unesco.org/webworld> [cited 2 May 2004].

much resemble the analogue original: the details and colour tones of a scene visible on a photographic print on paper can be digitised and projected on to a computer screen with no visible loss of quality.

Under the name 'American Memory' the Library of Congress was one of the first libraries in the world to build a digital library containing a wide range of historical materials.[2] In the period 1990-1995 a pilot project was carried out, followed by the development of an operational service. The system has been enhanced iteratively in the light of experience, as materials have presented new challenges and users have expressed new expectations. The long-term perspective has always been important, as is illustrated by the following quote: 'The resources created must serve for decades or centuries' ([ARM01] p. 46). One of the current challenges for American Memory is the facilitation of continuing access in the face of changing technology. The durability of the digital content is an important issue for the Library of Congress, and a clear solution for long-term preservation is still being debated.

Digital preservation is also an important subject in the authoritative reference book *Moving theory into practice. Digital imaging for libraries and archives* [KEN00]. The book gives an overview of policies and strategies towards digital conversion of archive and library material. Preservation of digital objects is also discussed. A recommendation concerning digital preservation in the reference book is the creation of metadata to support future preservation strategies ([KEN00] p. 143). This action line is an important issue in this study. Metadata is data or documentation about an object and describes various attributes. It gives the objects meaning, context and organisation.

Digitisation of cultural heritage by memory institutes is becoming more and more a supra-national issue. A number of international initiatives are playing an important role in the distribution and coordination of knowledge and experience concerning the digitisation of cultural heritage. Two of these initiatives are the Minerva network and the DigiCULT project. The aim of the European Minerva network is to discuss, correlate and harmonise digitisation of cultural and scientific content. The Minerva network aims to coordinate national programmes and its approach is strongly based on the principle of embedding national digitisation activities.[3] The *Technology Watch Reports* and *Thematic Issues* created by the EU-funded DigiCULT project cover a wide range of issues of great importance for the digitisation of the holdings of memory institutes.[4]

[2] The website of the American Memory digital library can be found at: <http://memory.loc.gov> [cited 2 May 2004].

[3] The website of the Minerva network can be found at: <http://www.minervaeurope.org> [cited 13 May 2004].

[4] The website of the DigiCULT project can be found at: <http://www.digicult.info> [cited 13 May 2004].

1.2 Longevity of digital objects

Among the first studies that addressed digital preservation in the cultural heritage sector were a special issue of the journal *History and Computing* devoted to the archiving of electronic records, published in 1992 (see: [ZWE92], [AND92], [DOO92], [THO92]), and the publication *Preserving the present* by Bikson and Frinking [BIK93], which was published in 1993. The *Scientific American* article *Ensuring the longevity of digital documents* by Rothenberg, published in 1995, is a widely cited article that started to raise a more general awareness of the problem that digital documents have a rather short life [ROT95]. Digital media 'will last forever – or five years. Whichever comes first' ([ROT95] p. 42). However, 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?

In the early 1960s the first social science digital data archives were founded, later followed by electronic text archives and historical data archives in the 1980s. The holdings of these data archives initially contained rectangular structured machine-readable files and these files are still accessible today. The durability of these data sets relies mainly on the encoding of the data in the ASCII data format and on the quality of the metadata connected to the datasets (ref. [DOO96]).

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 ([JON01] p. 10).

In 1998 Ross discussed the influence of digital preservation issues on the future of scholarship. He distinguishes three classes of digital materials – retroconversion, new digital content and by-products of contemporary life – that will form the digital record of the future. Scholars must be aware that active involvement in documentation issues of digital materials is essential for long-term access to them [ROS00]. Librarians and archivists must cooperate in order to tackle the risk that digital records will get lost.

Various studies take the preservation of digital objects into consideration. The studies by Dollar [DOL00], Jones and Beagrie [JON01] and Thibodeau [THI02] are among the most important publications in the field of the preservation of digital objects created by memory institutes. Most studies have a broad view on the type of the digital objects as subject for long-term preservation. The digital objects discussed in the studies range from single objects such as electronic documents to electronic records and extended computer programs. The literature distinguishes a wide range of ways to overcome technological obsolescence. Some proposed solutions exist only in theory and are not carried out in practical situations, or are carried out as 'proof of concept'.

By the year 2000 three main strategies towards digital preservation have been described: ([JON01] p. 26)

- *The technology preservation strategy*. Preservation of the original software and hardware that was used to create and access the information. This 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 yet resulted in a common ground on how to implement these strategies in memory institutes nor on which preservation strategy allies with what type of digital material. Currently a number of experiments and feasibility studies are being carried out. It can also be observed that, following the scientific data archives, memory institutes are implementing organisational structures that are committed to the preservation of digital data.

Another observation to be made is that the three digital preservation strategies mentioned above are applied for different purposes and user groups and to a wide range of digital materials such as computer programs, digital images, electronic texts and web pages. The background and perception of the people implementing the digital preservation strategies determine how the digital materials are actually understood and classified. Is a website a form of 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 being 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 will be able to judge whether the digital preservation strategies implemented today were the right ones. The aim of this study is to contribute to a better understanding and implementation of preservation of digital objects.

1.2.1 Storage media for digital objects

Naturally, the longevity of digital objects is determined by the stability of the medium on which the objects are stored. In this section the most relevant issues concerning the care and handling of storage media for digital objects are described.

In 1995 the US Department of Defense asked the National Media Laboratory to carry out 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 recorder used to write to the storage medium ([DOL00] p. 215). 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 Tape (DLT) and CD-ROM as optical disk. The best products have a life expectancy of at least 100 years. Assumed is that new media are used, that the media is accessed infrequently, that the media is consistently stored under the indicated environmental conditions and that the storage environment is clean and free of dust, smoke, food, mould, direct sunlight and gaseous contaminants.

A recent update on the state of the art concerning methods for the care and handling of optical storage can be found in [BEY03]. The report provides guidance on how to maximise the lifetime and usefulness of optical disks, specifically CD and DVD media, by minimising the risks of information loss caused by environmental influences or physical handling. An accelerated ageing study estimated the life expectancy of one type of DVD disk to be 30 years if stored at 25 degrees Celsius and 50% relative humidity. This testing is in the preliminary stages and much more needs to be done (ref. [BEY03] p. 13).

Despite the fact that paper and microfilm as a data storage medium generally have a longer life expectancy than optical and magnetic media, the durability of digital data expressed as collections of bits and bytes will be good enough for reliable storage for 100 years. The standard [ISO18921:2002] is available to estimate the life expectancy of CD-ROMs based on the effects of temperature and relative humidity. The purpose of this 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. A number of vendors apply this method and claim a life expectancy of about 200 years under optimal conditions.

Frey states: 'Since information technology is evolving rapidly, the lifetime of both software and hardware formats is generally less than the lifetime of the recording media' ([FRE01] p. 167). 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 of losing the digital data is posed by the fact that the interpretation and processing of the data require applications that can become obsolete. The durability of the data format is of greater importance than the durability of the storage medium.

1.2.2. Longevity of digital surrogates of historical photographs

At the moment, several digital image collections are threatened to become inaccessible because of the obsolescence of the image format, storage medium and the information system that provides access to the images.[5] A lot of effort is being put into the creation and dissemination of digital images of visual sources while long-term access issues are being neglected. Concerning image databases, the rapid changes in the computing industry are having two effects. First, users have higher expectations and systems place higher demands regarding digital images in terms of dynamic range and resolving power. Secondly, new file formats, new compression methods and new storage protocols can result in a situation in which legacy image databases can no longer be accessed. These new expectations regarding images and the evolving new technology (with the risk of older technology becoming obsolete) are a continuous threat.

Compared with textual sources, the digitisation of visual sources is more appealing because of its stronger impact on the general public. Although most humans are accustomed to pictures, they are still fascinated by them. One reason for this is the assumption that pictures do not lie. People believe what they see, because what they see cannot be false ([STR97] p. 6). Also, researchers in the Humanities field, such as historians, are increasingly discovering the value of visual sources for scientific research [CHO03]. Long-term access to durable, authentic and high-quality digital images is an important facilitator for the use of visual sources.

An enormous number of historical photographs have been converted into digital form, yet no fundamental research on the durability of this specific type of digital object has as yet been carried out. By limiting the scope of this study to that specific type of analogue source in memory institutes – historical photographs – this study will provide concrete solutions for durable digital surrogates of a specific type of visual source. Limiting the scope to this particular type of document makes it possible to create clear definitions and a common understanding. Nevertheless, the findings may have relevance to a wider range of digital objects and to other application areas, such as document management.

[5] On the occasion of the celebration of the fifth centenary of the voyage to America by Columbus, between 1990 and 1992 about 9 million archival documents were digitised at the Archivo General de Indias at Sevilla, Spain (see [GON92] and [GON99]). The project, a collaboration between the Spanish government, a private organisation and IBM, attracted a lot of attention in the 1990s and was considered as an important example of a good digitisation project for archives. In 2004 on the Internet no trace of the digital surrogates of the digital archive can be found or any recent reference to the digital archive.

A digital image is basically a raster where the points of intersection represent a colour, expressed as a computer code. These codes are created during the digital conversion process of the analogue original and stored in digital image files. Computer programs and hardware can convert the raster codes into coloured dots on a computer screen, translate these codes into instructions for other output devices such as printers or can manipulate the values of the codes, for instance, to improve the contrast of the scene visible on the image. The number of picture elements (or pixels) in a digital image and its colour gamut are the main distinctive features of a digital image. Often the quality of a digital image is determined by relating the physical characteristics of the original to the digital image. The most important quality aspects of a digital surrogate are the reproduction in the digital image of the tone scale, the image detail and colour of the original. The computer codes representing the pixels can be considered as a bitstream.

In the first instance, the digitisation of two-dimensional visual material, such as photographs, drawings, (photomechanical) prints and paintings, involves the conversion of a specific analogue storage medium into a digital format. Several guidelines are available for digitising visual sources, for instance, the *Guides to quality in visual resource imaging* [GUI00A]. Increasingly, official standards can be applied in parts of the digitisation process, such as the ISO standard [ISO3664:2000] to define the ideal viewing conditions of digital image files displayed on a computer monitor, independent of any form of hard copy. Reference [MRP04] contains a review of standards that are relevant for characterising the quality performance of digital cameras used to image cultural heritage.

From the early days of the application of digital imaging in memory institutes durability has been an important factor. During the pilot phase of the 'American Memory' project the notion was present that there is a difference between the creation of a digital surrogate for access on the one hand and the creation of a digital surrogate for preservation on the other hand [FLE92]. Most guidelines and best practice contain the advice to benchmark the digitisation specification. This benchmark should be based on an assessment of the intended use of the digital surrogates, the characteristics of the collection and the available resources. As it is impossible to know the requirements of future generations regarding images, most digitisation specifications are based on well-defined short-term usage requirements. The image quality must be high in order to enable a number of types of usage, even in the distant future. Next in importance, the documentation of the digital images must be sufficiently rich and detailed for future generations to understand the specifications and context of the bitstream that represents a digital image. This implies that the user can understand the syntax and semantics of this bitstream. One of the main subjects of this study is the creation of metadata required to guarantee long-term access of digital surrogates of historical photographs. Both the syntax and semantics of the metadata are covered.

Memory institutes consider preservation of digital assets as an important issue, but in the first instance the short-term use of the digital objects is the main rationale for carrying out a digitisation project. Most of the time ‘access’ is the main purpose for setting up a digitisation project. By creating digital surrogates of original objects, access to the collection can be improved dramatically. In the event that users do not need to gain access to the original – or are not allowed to have access to the original – preservation of the original is facilitated by digitisation as well. This is a form of passive preservation of the original. The digital surrogate prevents the decay of vulnerable objects simply because the latter are not touched any more. So, in the first instance, the specifications of the digital surrogates are based on the short-term usage of the digital objects. Attention is given to the preservation of these digital surrogates. As digitisation is a fairly expensive activity, it is considered important to avoid the risk that the conversion has to be repeated in the future. Another reason for the relevance of preservation is the fact that a digital object can be used in the future for purposes other than its original purpose.

Despite the fact that the importance of digital preservation is apparent for memory institutes, the objects to be preserved are not durable by definition. Often the recommended image file formats for access are compressed to decrease the file size and improve the transmission speed via networks. It is a good practice to derive these compressed images from uncompressed, un-processed, ‘raw’ master files and to consider these master files as the digital objects to be preserved for the long term. As new image file formats appear, the requirements for archival image file formats can change over time. It is necessary to monitor actively a digital master file and apply a preservation strategy in order to keep the master images vital. Both the pixels representing the digital image and metadata containing their relevant documentation are required for long-term access to the digital objects.

1.3 Research

This study takes the durability of a specific digital object into consideration, namely digital surrogates of historical photographs. The wide range of types of digital objects and the rapid changes in information technology make examination of the durability of digital objects in general unfeasible, whereas the in-depth examination of a clearly defined digital object is more realistic.

Another reason for the concentration on a specific type of digital object is that it is assumed that the emphasis on a digital object of which detailed features are known may result in more effective digital preservation solutions. An important specific quality of digital surrogates of historical photographs is the close relation between the analogue original and its digital representation with respect to the details of the scene depicted on the image. This aspect plays an important role in the digital preservation approach of memory institutes.

The purpose of this research is to illustrate a means for memory institutes to create durable digital surrogates of photographs, to manage them and to provide long-term access to them. The research on the digital preservation of this very specific digital object will also contribute to better insight into the longevity of digital objects in general. Some of the results of this research may be relevant for the preservation of other types of digital objects.

1.3.1 Research question

Concerning quality issues related to the digitisation of analogue photographs, a number of publications and studies are available. The works by Frey and Reilly [FRE99] and the *Guides to quality in visual resource imaging* [GUI00A] are among the most important studies in this field. Quality digitisation implies that the significant features of the analogue original are available in the digital surrogate. For this, a thorough assessment of the analogue original is required as well as a benchmarked digital capture process. In the literature, the two aspects – the longevity of a specific type of digital object and the benchmarked digital capture of historical photographs – are not addressed as an integrated issue in a detailed way. In order to address this issue the following research question is formulated:

> How can benchmarked digital surrogates of historical photographs be preserved?

In order to answer this question the following research goal has to be achieved:

> Identification and assessment of relevant building blocks that enable the creation, management and long-term access of benchmarked digital surrogates of historical photographs.

This study offers both practical and theoretical contributions. The practical contribution of the research is that it offers memory institutes components that can be used to improve the durability of digital objects. This research contributes to theory by combining previously unrelated studies about digital conversion of analogue sources and studies about digital longevity, and extending them into an integrated approach.

1.3.2 Research approach

The research approach concerns the activities that are carried out in order to achieve the research goal. The purpose of this research is twofold. The first purpose is to describe the problem of long-term access to digital objects in general and long-term access of digital surrogates of historical photographs in particular. Long-term is defined as long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community ([ISO14721:2003] p. 1-1). The second purpose of this research is to provide memory institutes with practical solutions for the creation, management and access of durable digital surrogates of historical photographs. An important aspect of the research approach is that, for the realisation of the practical solutions, existing building blocks are used. Building blocks can be defined as procedures, tools, specifications, standards and guidelines available to realise the creation of digital durable surrogates of historical photographs [HOR02].

A research approach consists of three elements ([VRE95] pp. 21-28):

- *Research philosophy*. A research philosophy underlines the way in which the data on the phenomenon studied are collected and analysed. It determines what kind of knowledge can be obtained and the limits of that knowledge.
- *Research strategy*. The research strategy concerns the steps that are carried out to execute the enquiry into the phenomenon studied.
- *Research instruments*. Research instruments are the tools to carry out or implement the research strategy.

1.3.3 Research philosophy

Longevity of digital objects is a relatively young field of research, applied in a fairly broad manner by a wide range of initiatives. The nature of knowledge, or epistemology, of this research subject does not have a tradition or 'school of thought'.[6] The determination of a research philosophy for this research must take this situation into account. A number of scientific disciplines can be distinguished as relevant for the research. This makes it difficult to determine whether an existing research philosophy used in a specific scientific discipline can be used to achieve the research goal. A number of scientific disciplines relevant for digital preservation are discussed. Examples of research topics on the scientific disciplines that are related to this research are given in Table 1.1.

The first contributing scientific discipline is the work related to the capture and creation of images, both analogue and digital. Jacobsen and Attridge [JAC00] provide a good overview of the state of the art of the techniques and technologies of photography, both in its analogue and digital form. Secondly, input from computer science (ref. [TAN98] pp. 310-311), the branch of engineering science that studies computable processes and structures, is relevant as far as it deals with the processing and storage of digital images. Thirdly, information science, defined as science concerned with the gathering, manipulation, classification, storage and retrieval of recorded knowledge, is relevant for this study. Digital preservation has a connotation of archival storage, making archival science also a contributing scientific discipline for this research. The archival perspective brings an evidence-based approach to the management of recorded knowledge [GIL00]. Contextual issues such as data integrity and object authenticity are important pillars under archival science. As scholarly use of digital objects created by memory institutes is often mentioned as important, the Humanities can be considered as a fifth contributing scientific discipline for this research.

[6] A number of potential frameworks for research on digital preservation are described by Hedstrom in:[HED02].

Table 1.1 Contributing scientific disciplines relevant for this research and some exemplary research topic

Scientific discipline	Exemplary research topic relevant for preservation of digital surrogates of historical photographs
Imaging science	Creation of benchmarked digital images
Computer science	Durable storage of digital surrogates
Information science	Creation and structuring of preservation metadata
Archival science	Object authenticity and object integrity
Humanities	Application of digital visual sources in scholarly research

An approach for constructing a research philosophy is to review methods to acquire scientific knowledge. Van Dooren mentions seven important directions for obtaining scientific knowledge of the outside world ([DOR92] pp. 46-49):

- *Realism*. This is a common-sense approach to a phenomenon as exposed to the researcher. An opinion on the phenomenon studied is the basis for the common-sense argumentation.
- *Empiricism*. Only sensorial experience can lead to knowledge. The difference between empiricism and realism is that empiricism is based on experience and not on opinions. The arrangement of the sensorial experience by the mind is an essential part of empiricism. The view is traced to Aristotle.
- *Rationalism*. Reason (based on the Latin 'ratio') is the source of scientific knowledge. Mathematical methods are applied to achieve empirical knowledge on a phenomenon. Rationalism is traced to Plato.
- *Idealism*. Experience is completely turned down as a source of knowledge. The 'ideas' behind the reason are the reality, because the outside world is created by reason.
- *Criticism*. We can only have certain knowledge of the outside world as far as it is revealed in a given situation. The philosopher Kant claims that 'the object of our knowledge is to be taken in a twofold sense: namely as appearance and as a thing in itself and objects are known to us only in the first sense, as appearances' ([GAR98] p. 578).
- *Pragmatism*. Acquisition of scientific knowledge is a practical activity and only valuable if it works in reality. Knowledge is an instrument needed for taking action. This approach is also called 'instrumentalism'.
- *Scepticism*. Based on the utterance of doubt that scientific knowledge is possible. Scepticism often emerges as a correction to a vision that is too self-assured. In classical antiquity, sceptics were 'those who suspend judgement'.

The application of any one of the knowledge acquisition directions guides the choice of research instruments with which data on the research subject are collected. Knowledge acquisition methods relevant for the research in this study are realism, empiricism, criticism and pragmatism. Realism requires common-sense argumentation and this is essential for solutions to preservation problems in an unknown future. Empiricism is important as an approach because of the arrangement of available experiences related to the research problem. The critical, Kantian direction allows induction based on a set of 'a priori sciences'. Pragmatism is applied in this research because available tools, procedures and guidelines relevant for the research question are taken into consideration.

A fundamental philosophical issue to be clarified is whether the research activities interfere with the problem being studied. In the literature, this distinction is described in a number of ways. De Vreede ([VRE96] pp. 21-22) makes a distinction between positivism as a philosophy that observes and describes reality in an objective way, without interfering with the phenomenon being studied, as opposed to interpretivism as a philosophy that understands reality. The interpretive research philosophy is based on subject-dependent perceptions and interpretations of it.

There is no lack of strategic views or theories on digital preservation. A number of principles or sets of statements are around, devised to explain how digital preservation can be realised by subjectively interpreting observations of reality. Van Dooren ([DOR92] p. 57) uses the conflict of methods in sociology to illustrate the difference between objective interpretation in scientific research and subjective interpretation. The critical–rational method requires a completely objective position of the researcher in order to establish a theory that solves the research problem. On the other hand, the dialectical method states that contrasts and interactions with the research problem result in a better solution.

This research investigates how long-term access and usage of a specific digital data object can be realised. An attempt is made to formulate sound scientific solutions that are applicable in the future, based on knowledge of the present and the past. Induction is the process of deriving general principles valid in the future based on particular facts or instances from the present and the past. Digital longevity has a strong time connotation, as only in the unknown future will it become clear whether the applied durability methods were appropriate. The research strategy has to make this unknown future as probable as possible.

It will be difficult, if not impossible, to answer the research question and obtain the research goal without determining the value of the rather large number of existing practices, strategies and theories that can be used to enable the durability of digital objects. By definition, this value determination is subjective, thus following an interpretive research philosophy. The acknowledgement that subjective reasoning is inevitable does not mean that it is not important to strive towards inductively strong conclusions. As induction is based on the interpretation of observations, it is important to assess the evidence of a prediction that is based on induction. 'It must be possible to decide whether someone is an expert predicator or a charlatan' ([SKY66] p. 53). The next section examines to what extent the research strategy can be based on inductive reasoning.

1.3.4 Research strategy

The research strategy concerns the steps that are carried out to execute the enquiry into the phenomenon studied. It outlines the sequence of data acquisition and analysis ([VRE95] p. 22). The research strategy is based on the nature of the research problem and on the status of the theory development in the research field.

Concerning digital preservation the three main strategies to prevent the obsolescence of digital objects are technology preservation, technology emulation and information migration (see section 1.2). There are a number of variants and extensions on these strategies. All digital preservation strategies are based on a number of general theories, premises and assumptions. Theories are mental constructions that shape the way we conceive. The truths that they yield are not objective, but must be seen in the light of the theory ([HOM04] p. 8). Popper and Lakatos acknowledge that theories precede observation ([CHA99] p. 156). Lakatos proposes to use the concept of 'research programme' for the theoretical framework of scientific research in empirical sciences. The research programme consists of methodological rules: some tell us what paths of research to avoid (negative heuristic) and others what paths to pursue (positive heuristic) ([CHA99] p. 158). Even science as a whole can be regarded as a huge research programme.

Induction and probability

Based on existing digital preservation strategies, this study develops and evaluates building blocks on the durability of digital surrogates of historical photographs. This implies an inductive research strategy: the shape of the future state of affairs is based on facts that are already known. One of the most important uses of inductive logic is to frame our expectations of the future on the basis of our knowledge of the past and the present ([SKY66] p. 20). Inductive logic is linked with the concept of probability. That is the likelihood that phenomena based on inductive reasoning turn out to be sensible in the future. The relation between probability and induction is very well explained by Skyrms [SKY66]. Correct reasoning and using compelling arguments are important parts of the inductive research strategy. An argument is defined by Skyrms ([SKY66] p. 2) as a list of statements, one of which is designated as the conclusion and the rest of which are designated as premises. An argument is evaluated on two grounds: Are the premises true? And – supposing the premises are true – what sort of support do they give to the conclusion?

Often only domain experts are able to judge the strength of the evidential link between the premises and the conclusion. If the premises, for instance, claim that a certain mechanism enables the faithful digital capture of colour data, one would ask an image scientist whether the claims are true. If the premises provide good evidence for the conclusion, the argument is said to be inductively strong. This concept is defined by Skyrms in the following way. An argument is inductively strong if, and only if, it is improbable that its conclusion is false while its premises are true. The degree of inductive strength depends on how improbable it is that the conclusion is false while the premises are true ([SKY66] p. 7). The type of probability that grades the inductive strength of arguments does not depend on the premises alone or on the conclusion alone, but on the evidential relation between the premises and the conclusion.

An example of an inductively strong argument consists of the three premises 'digital data stored on computer tapes is still readable today', 'digital data stored on optical media is still readable today' and 'digital data stored on floppy disks is still readable today'. Based on these three premises, the conclusion 'digital data stored on hard disks is still readable today' can be formed. This conclusion is not in itself probable. But it is improbable that the conclusion is false given that the premises are true. Whether the premises stated in the example are true is another issue, not discussed further at this point.

The induction problem

The induction problem concerns the impossibility of deriving a universal conclusion based on a limited number of observations. The sole fact that yesterday the sun came up does not guarantee that it will happen tomorrow. According to Holland and others, induction has been called the 'scandal of philosophy' ([HOL89] p. 1).

The inductive probability of an argument is a measure of the strength of the evidence that the premises provide for the conclusion. In a lot of cases intuition plays a role in evaluating the probability of an argument. When we state, for instance, that it is improbable that data stored on punch cards is still usable today we rely on some intuitive sense of probability stating that punch card readers are no longer available. This intuitive sense of probability is called epistemic probability by Skyrms. The epistemic probability of a statement is the inductive probability of that argument which has the statement in question as its conclusion and whose premises contain all of our relevant factual knowledge ([SKY66] p. 15). The epistemic probability of a statement can vary from person to person and from time to time. In principle, it is possible that punch cards are still used as a storage medium for data, but within the framework of this study the epistemic probability of this statement is quite low. The epistemic probability of a given statement can also change over time for a particular person, because human knowledge is continually in a dynamic process of simultaneous growth and decay.

There are no universally accepted rules for constructing inductively strong arguments. It is also problematic to measure the inductive probability of an argument. A system is required that accords well with both scientific practice and common sense, and that is precise and clear. Skyrms calls this system 'scientific inductive logic'. Scientific inductive logic classifies inductively strong arguments as having a high inductive probability.

Scientific inductive logic

Before a system of scientific inductive logic can be constructed, Skyrms discusses the rational justification of the use of scientific inductive logic. Skyrms suggests that a system of inductive logic is rationally justified if it can be shown to be an embodiment of those inductive rules of science and common sense that we take to be a standard of rationality ([SKY66] p. 48). Being still a very weak rational justification, Skyrms next tries to construct a system of scientific inductive logic. Arguments that give high probability rather than certainty are still good arguments. It is difficult to gain certainty in a world of change.

A system of inductive logic is rationally justified if the arguments yield true conclusions from true premises most of the time. This justification is described and criticised for the first time by the 18th century Scottish philosopher David Hume. Formulating the rules for inductive logic appears to be more difficult than doing the same for deductive logic. Skyrms states that deductive logic is a 'yes or no' affair. An argument is either deductively valid or it is not. While deductive logic must classify arguments as valid or not, inductive logic must measure the inductive strength of arguments ([SKY66] p. 52). Several philosophers are convinced that it is not possible to construct a system of scientific induction. Prediction of the future is an art, not a science. We must rely on the intuition of experts rather than on scientific inductive logic to predict the future. Whether a certain situation involves change or not may depend on the descriptive machinery of the language used to discuss that situation.

This is demonstrated by the Goodman paradox ([SKY66] pp. 57-69) and ([HOL89] pp. 234-235). Goodman invites us to consider a new colour word: 'grue'. Something is grue if it has the colour green *before* a given point of time and something will be grue if it has the colour blue *during* or *after* this point of time. Suppose this point of time is the year 2000. This would mean that a green grasshopper before the year 2000 and a blue sky during and after the year 2000 would both have the same colour name. Goodman also throws in the colour word 'bleen', which applies to anything that looks blue before a given point in time and green afterwards. How do we know that the grass is green and not grue before the given point in time and that the sky is blue and not bleen? Suppose that a person speaks the 'grue' and 'bleen' language. This means that predictions about the colour of objects (made before the year 2000), based on projections into the future, can be demonstrably wrong: the colour of the same object has changed from green to blue (Skyrms uses the colour of emeralds as a real world example).

The Goodman paradox demonstrates how difficult it is to assess the inductive strength of an argument. The 'grue' and 'bleen' discussion makes the following clear ([SKY66] p. 66):

- Whether we find change or not in a certain situation may depend on the linguistic machinery we use to describe that situation.
- What regularities we find in a sequence of occurrences may depend on the linguistic machinery used to describe that sequence.

- We may find two regularities in a sequence of occurrences, one projectable and one unprojectable, such that the predictions that arise from projecting them are both in conflict.[7]

In line with Skyrms, it can be concluded that projecting observed regularities into the future is not as simple as it appears in the first instance. The regularities found in a certain sequence of events may depend on the language used to describe that sequence of events. The Goodman paradox showed that, if we try to project all regularities that can be found by using any language, our predictions might conflict with one another. There is a need for rules for determining projectability in scientific induction ([SKY66] p. 61). The problem of formulating these rules is called the new riddle of induction. Skyrms states that solutions for this new riddle of induction have not yet been found. He mentions some 'building blocks' that can be used to fill the gap between intuition and a complete system of inductive logic, based mainly on mathematical theories of probability.

Conceptual spaces

Gärdenfors [GAR00] studies human inductive competence and develops a theory of constraints for inductive inferences. In line with Skyrms, the focus is on projectability, that is, which properties and concepts may be used in an inductive research strategy. Gärdenfors concentrates on the way humans observe phenomena insofar that these observations are the basis for inductive reasoning. Three levels of accounting for observations are distinguished ([GAR00] p. 204):

- *The symbolic level.* This way of representing observations consists of describing them in some specified language. The language is assumed to be equipped with a fixed set of primitive predicates and the denotations of these predicates are taken to be known.
- *The conceptual level.* The observations are not defined in relation to some language but characterised by an underlying conceptual space. Induction is seen as closely related to concept formation.
- *The subconceptual level.* Observations are characterised by inputs from sensory receptors. The inductive process is conceived as establishing connections among various types of inputs. A popular way of modelling this kind of process is to use artificial neuron networks.

The three levels are discussed briefly. Gärdenfors argues that, depending on which of the three approaches mentioned above is adopted, thoroughly different considerations concerning inductive inferences will be brought into focus. The

[7] The higher the degree of inductive strength of an argument, the more projectable it is. An example of a projectable argument: If one hundred per cent of observed samples of pure water have had a freezing point of 0 degrees Celsius, the next observed sample of pure water will also have a freezing point of 0 degrees Celsius. An example of an unprojectable argument: If one hundred per cent of the recorded economic depressions have occurred at the same time as large sunspots, the next economic depression will occur at the same time as a large sunspot (ref. [SKY66] pp. 55-56).

symbolic level is strongly connected to logical positivism, which led to serious problems in relation to induction. The problem of projectable inductions is basically a problem of representing information, as is illustrated by the Goodman paradox. Gärdenfors concludes that the symbolic approach to induction sustains no creative inductions, no genuinely new knowledge and no conceptual discoveries ([GAR00] p. 211).

The conceptual level of inductive reasoning involves the establishment of connections among concepts or properties from different domains. On the conceptual level, Gärdenfors proposes a solution to the Goodman paradox. An observation is defined as 'an assignment to an object of a location in a conceptual space' ([GAR00] p. 211). What counts as a natural property depends on the underlying conceptual space. Given the standard representations of colours as a point in a colour space, 'green' and 'blue' are natural properties while 'grue' and 'bleen' are not, because they presume two dimensions, colour and time, for their description. The subconceptual level concerns inductive processes below the conceptual level. Humans have powerful abilities to detect multiple correlations among different domains. Gärdenfors discusses a number of ways in which machines are equipped with sensors. The subconceptual induction process merely involves the interpretation of the uninterpreted data by artificial intelligence techniques and is not relevant for the inductive research strategy of this study.

Gärdenfors states that, if inductive reasoning is studied on the conceptual level of representation instead of on the symbolic level, the classical riddles on induction can be circumvented ([GAR00] p. 3). Thus, it seems that a theory on conceptual spaces is the nearest one can get to a solution for the induction problem. A theory of conceptual spaces is a particular framework for representing information on the conceptual level. A conceptual space is built upon geometrical structures based on a number of quality dimensions ([GAR00] p. 2). Conceptual spaces are theoretical entities that can be used to explain and predict various empirical phenomena relating to concept formation.

The epistemological role of the theory of conceptual spaces is to serve as a tool in modelling various relations among our experiences, that is, what we perceive, remember or imagine. Concepts are not independent of each other but can be structured into domains; for instance, spatial concepts belong to one domain, concepts of colour to a different domain, concepts of sounds to a third and so on. Fundamental is to build up domains for representing concepts. The structure of many quality dimensions of a conceptual space will make it possible to talk about distances along the dimensions. The smaller the distances between the representations of two objects, the more similar they are and the stronger the inductive strength of a statement will be that is based on this conceptual space. Some easy to understand quality dimensions are those closely connected with what is produced by our sensory receptors, such as sight, temperature and weight. There is, however, a wealth of quality dimensions that are of an abstract non-sensory character and digital durability is definitely an example of such a quality dimension. Culture, in the form of interactions among people, may in itself generate constraints on conceptual spaces.

In [HOL89] the induction problem is resolved in a less abstract manner than is the case with Gärdenfors. Induction is studied in a more pragmatic context by insisting that sensible inferential rules take into account the kinds of things being reasoned about. By stating, for instance, that 'grue' (ref. Goodman paradox) is not a pragmatically useful category the paradox can be resolved ([HOL89] p. 7). Induction is considered as highly context dependent as it is being guided by prior knowledge activated in particular situations.

The inductive research strategy

The nature of the research question of this study allies with the inductive research strategy. As the induction problem has not been solved, relevant expert knowledge is the main factor in the predictive strength of the arguments that are used to answer the research question. The use of conceptual spaces as defined by Gärdenfors is one of the nearest approaches one can get to the solution of the induction problem. The conceptual level of inductive reasoning involves the establishment of connections among concepts or properties from different domains. The conceptual space for this research is based on connections between expert knowledge originating in a number of scientific disciplines, as illustrated in Table 1.1. Inductively strong arguments are based on correct reasoning and on using compelling arguments. Natural language is the main vehicle to construct the inductive arguments used in this research. The research strategy used in this work consists of six activities:

1. Current practices and existing theories relevant for the research problem are identified and described.
2. The conceptual space is formulated by identifying essential aspects related to a number of contributing scientific disciplines. This activity can be considered as the epistemological point of departure (chapter 2).
3. By using inductive reasoning a number of premises are constructed that contain relevant factual knowledge (chapter 3).
4. The validity of premises is examined by carrying out 'experiments'. The experiments in this case consist of the usage and evaluation of available procedures, tools, specifications, standards and guidelines. Thus, only existing building blocks are used (chapter 4).
5. The experiments are evaluated. Additional requirements for improving the building blocks for the durability of digital surrogates of historical photographs may be identified (chapter 5).
6. The premises and experiments result in concluding remarks that are related to the point of departure of this study (chapter 6). The conclusions will be relevant for the longevity of digital surrogates of historical photographs and are broadened for the longevity of related types of digital objects in general.

2.3.1 Research instruments

Research instruments are the tools to carry out the research strategy. Chilvers is one of the few scholars in the field of digital preservation research with a clear view on appropriate research instruments. In her research on the long-term access of digital data objects she considers the Soft Systems Methodology (SSM) as an appropriate means to use ([CHI01] p. 153).[8] Chilvers examines the reasons why existing management practices appear to be inadequate for managing long-term access to digital data objects. SSM helps to analyse the management of digital data objects in conjunction with social context. Chilvers characterises the digital preservation research area as an 'ill-defined people-centred problem situation' ([CHI01] p. 152). The main result of Chilvers' research is the recommendation to develop a 'super-metadata framework', designed to create a supportive structure to allow for past, present and future metadata developments within the information community worldwide. The relevance and application of metadata constructs is also an important issue in her dissertation.

The SSM method is not applicable as a research tool for this research, because the management aspects of digital preservation are less apparent than technical issues. Qualitative data resources are the most important resources used in this study, such as literature, reports and personal observations.

Two EU-funded projects serve as important foundations for the research instruments related to the research strategy of this work. These are the EVA project [HOR01A] and the SEPIA project [LUS02]. The main goal of the EVA project ('European Visual Archive') was to investigate the creation of digital surrogates of historical photographs that are part of the holdings of public archives, and the development of an information system to provide access to the digital surrogates.[9] The main goal of the SEPIA ('Safeguarding European Photographic Images for Access') project was to improve the preservation of photographic materials. Both projects meet the criteria for case studies as defined by De Vreede ([VRE95] p. 28): the researcher is an observer, focusing on 'how' and 'why' questions.

Building blocks

The most important instrument for carrying out the research in this study is the description, assessment and application of building blocks, defined as existing procedures, tools, specifications, standards and guidelines available to realise durable digital surrogates of historical photographs. The functions of the building blocks are to enable the creation, management and long-term access of the digital surrogates. The rationale behind the use of the concept of building blocks is that it is assumed that (re)use of existing knowledge is much more efficient than develop-

[8] SSM is a way of dealing with problem situations in which there is a high social, political and human activity component.

[9] The EVA project was succeeded by the project EVAMP (EVA Market Validation) whose main goal is to create a business plan for the results of the EVA project.

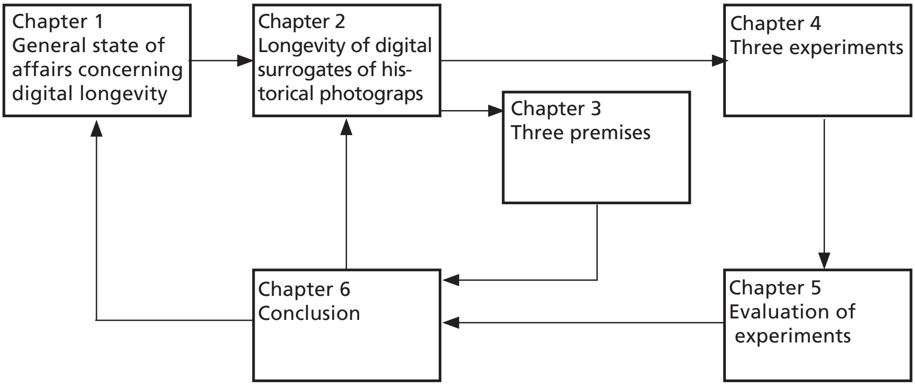

Chapter 1
General state of affairs concerning digital longevity
Chapter 2
Longevity of digital surrogates of historical photograps
Chapter 3
Three premises
Chapter 4
Three experiments
Chapter 5
Evaluation of experiments
Chapter 6
Conclusion

Figure 1.1 Outline of the study

ing knowledge from scratch. Moreover, it is assumed that users will be inclined to accept existing, assessed, mature solutions more easily than new suggestions.

It should be stressed that the building block concept is very applicable for the problem area of digital preservation, because cooperation and consultation is a common activity for the stakeholders in this field. Cooperation and consultation activities are in line with the proposed building block approach in this study. The research is based on an analysis of existing building blocks that can be used to apply or test theories relevant for the durability of digital surrogates of historical photographs. Examples of building blocks are a software tool to read or write embedded preservation metadata in a digital image, a guideline to create benchmarked digital surrogates, or a specification of the use of the XML data format. The application of the tools can be considered as experiments and as such be appointed as research instruments in this research. The Internet provides the opportunity to find a huge number of existing tools, specifications, standards and guidelines, making it superfluous to develop new, dedicated building blocks.

2.3.2 Research outline

This study consists of six chapters. The structure is illustrated in Figure 1.1. The first introductory chapter contains the motivation, problem statement, scope and research approach of the research. In the first chapter also the current state of affairs concerning the application of digitisation in memory institutes is described as well as the main strategies relating to digital preservation.

The second chapter, 'Digitisation and digital longevity in memory institutes', contains the specialist knowledge needed to understand the work in this study and the knowledge upon which the work is built. The two main aspects of digitisation of analogue sources are the conversion of the analogue medium and the creation of documentation on objects in digital format. The first part of chapter 2 describes why and how memory institutes convert historical photographs into digital surrogates. The second part of this chapter describes the digital longevity landscape, as it is part of the current daily practice within memory institutes.

The third chapter, 'Durable digital surrogates of historical photographs', presents three premises that are relevant for the durability of digital surrogates of historical photographs, based on theories and practices derived from existing knowledge not necessarily related to digital longevity. First, the benchmarked digitisation of analogue originals is covered. The second part of chapter 3 deals with preservation metadata, and the third part of the chapter concerns the durable access and persistent storage of the digital objects.

Chapter 4, 'Experiments', consists of a validation of three hypotheses upon which the durability of digital objects can be based and demonstrates how the methods can be used in practice. Attempts are undertaken to apply theories that are considered to enable the longevity of digital objects. These three theories are the theory of applying image file format standards in order to realise digital longevity, the theory of applying the XML data format in order to realise longevity, and the theory of creating standardised preservation metadata in order to realise longevity.

Chapter 5, 'Evaluation of the experiments', consists of the confirmation or repudiation of the theories discussed in chapter 4. Arguments are given that either strengthen or contradict the outcomes of the experiments.

Chapter 6 contains the conclusions of the research. The chapter contains an extrapolation of the outcomes of the experiments. Also, the premises discussed in chapter 3 are taken into consideration. This implies the estimation of generic solutions for the durability of digital objects in general and of digital surrogates of historical photographs in particular.

2 Digitisation and digital longevity in memory institutes

The research problem of this study addresses the threatening loss in the future of digital surrogates of historical photographs created by memory institutes. This chapter contains the context of the research problem: the specialist knowledge needed to understand the work in this research. Digital conversion in principle implies the translation of the physical structure (for instance, the silver grains of a photographic print) into digital coded picture elements, or pixels. As the content in digital form is not fixed to a specific output medium – as is the case with a 'traditional' analogue source – the representation of the document content is determined by the characteristics of the image output device (for instance, a computer screen or paper).

Compared with textual documents it is much more difficult to digitise graphical, non-textual sources. The main reason for this is that for the full informational capture of textual material an objective benchmark method is available ([KEN96] pp. 12-26), while for non-textual material no reliable quality assessment method is available to date. A corpus of textual material contains items with similar features, whereas the features of items in a corpus of graphical material are much more heterogeneous. This means that within a collection of analogue graphical sources the digitisation settings might be changed, whereas within a collection of textual material the same digitisation settings can be used for all individual items of a corpus.

A major characteristic of digitisation projects of graphical sources in the cultural sector is the aim to create so-called digital master files, rich enough to be suitable for future use. It is obvious that in the first instance the digital images will be used for faithful on-screen viewing or the creation of a reproduction on paper. As the output requirements may change in the future, it is evident that the quality and longevity of the digital objects are important. Instead of a once-only activity the digitisation of graphical sources requires more continuous attention. An active archiving and access policy is required in order to facilitate specific future utilisation, such as scholarly research or online viewing. The concept 'use-neutral digital master image' is used to express the requirements of high-quality digital objects created in a conversion project involving analogue originals. With this approach, an original is digitised once, at the highest level of quality affordable, and studio standards such as colour matching and contrast levels are set so that the master image can be used for multiple applications.

This chapter consists of two parts. First, the digitisation of photographic collections by memory institutes is surveyed. The second part of the chapter contains a discussion of how memory institutes implement the digital longevity of their digital assets.

2.1 Digitisation of historical photographs

This section contains an overview of the way libraries, archives and museums convert historical photograph collections into digital form. It consists of six parts. First, the motives for digitising historical photographs are described. The second part of the section contains a classification of source documents. This classification is required because the physical features of sources influence the digitisation process. In the third part of the section photographs as a specific type of document are described in more detail. Part four of this section contains information on the digital capture of graphical sources. The fifth part of this chapter section covers graphics file formats that enable the storage of picture elements or pixels. The last part deals with digital master files. These digital objects are the digital surrogates that contain all the significant features of the analogue original.

2.1.1 Why digitise?

The two main reasons for memory institutes to digitise photographic items from their holdings are to improve access to the items and to preserve the vulnerable originals. Preservation implies the protection of originals by delivering digital surrogates to users. A vulnerable original no longer needs to be touched. This is a form of 'passive conservation' as long-term keeping of unstable photographic originals is only possible with cold storage [FREY97B], [BAT03].

The second reason to digitise items is the improved access to the intellectual content of the items. The Internet can facilitate transparent and easy access to digital surrogates of analogue sources. Digital data can be copied without loss and it is easy to share digital information with a number of users. These arguments for digitisation of photographic collections are supported by a survey carried out in 1999 (see: [KLI00] p. 25). The report *Digitising historical pictorial collections for the Internet*, published in 1998, contains the following quote, which describes very well the then existing situation:

> Historical photograph collections are among the least accessible sources available to researchers because of their large size, complex organisation, physical fragility, and often-rudimentary description and cataloguing. Most consist of large groups of related materials that share one or more significant common denominators, such as source, subject, or medium. That common feature often serves as the framework for organizing and providing access to the individual pieces ([OST98] p. 8).

Since then the situation has improved considerably. More attention is being given to the preservation of historical photographs and the access to photographic items has also improved. The two main reasons for this are the increased attention to photography in general and the greater availability of digital capture and presentation techniques. Both factors are clarified.

Officially, photography was invented in 1839 and its 150th birthday turned out to be an influential catalyst for the emancipation of photography as an established form of art. Publications on the history of photography, such as *A new history of photography* [FRI98] and *The art of fixing a shadow* [GRE89] raised the awareness that photography can be considered as a distinct form of art. Another indicator of the growing attention to photography is the compilation of specific collections, such as the National Photo Collection in the Netherlands in 1994 [BOO96]. In the slipstream of the growing interest for fine-art photography, created by famous photographers or otherwise recognised by experts, the attention for archival photograph collections has also increased. The distinction between fine-art photographs and historical, documentary photographs is not always clear, as historical collections can contain important, high-quality items. In general it can be stated that a fine-art photograph collection consists of a relatively small number of valuable items, whereas an historical photograph collection contains a relatively large number of items whose main function is to document events and objects.

The second reason for the increased interest in visual sources are the possibilities for conversion and access provided by information technology. The promising prospects of the potential of the Web were the main force behind the large number of initiatives to create digital cultural heritage document repositories.

A number of publications have influenced the way memory institutes started to apply digitisation techniques. In 1996 the Getty Art History Information Program published *Introduction to imaging* [BES96], introducing digital imaging technology and vocabulary as they relate to the development of image databases and outlining the areas in which institutional strategies regarding the use of imaging technologies must be developed. The publication was aimed mainly at the museum community. Also in 1996 *Digital imaging for libraries and archives* [KEN96] was published, succeeded in 2000 by *Moving theory into practice. Digital imaging for libraries and archives* [KEN00]. This book is intended as a 'self-help reference for libraries and archives that choose to retrospectively convert cultural resources to digital form'. The *Handbook for digital projects: a management tool for preservation and access*, published in 2000 [SIT00], is another influential publication used by memory institutes. In the United Kingdom the 'Arts and Humanities Data Service' (AHDS) provides memory institutes with publications and other services on digitisation, and in the Netherlands the project 'Geheugen van Nederland' (Memory of The Netherlands) assists memory institutes with guidelines on the conversion of a wide range of collections, including photographs.[10]

[10] The website of the Arts and Humanities Data Service is <http://www.ahds.uk> [cited 15

Two publications are particularly relevant for the digitisation of photographs. The first is *Digital imaging for photographic collections. Foundations for technical standards* [FRE99]. This book tries to answer the question to what extent it is possible to capture digitally all the information in photographic originals. For this, an adequate capture of detail and contrast of the originals is required and the book explains what kinds of measurements are relevant. Specific image science knowledge is required in order to apply the methods described in [FRE99]. When this book was published (in 1999), several standards were still under development, but since then the situation has improved. Section 3.1 of this study contains a review of the state of the art of standards available to create high-quality digital surrogates of analogue photographs. The second publication that contains important information on the digitisation of photographic collections is *Guides to quality in visual resource imaging* [GUI00A].[11] This guide, published in 2000 and only available online, covers topics like 'selecting a scanner', 'the range of factors affecting image quality' and 'file formats for digital masters'.

Digitisation implies two kinds of activities: medium conversion and creation of metadata in electronic form. Medium conversion involves the creation of a digital version of a document: a digital image or bitmap. Creation of metadata involves the creation of digital documentation that enables functions such as the retrieval, administration, storage and durability of documents in either digital or analogue form.

In the first instance, digital image collections are created in a project-oriented environment with a limited duration and scope. Typically the main scope of a project is to have a collection converted into digital form, often without having a specific idea of the intended usage of the images. To gain hands-on experience with new technology is an important reason to convert a collection. After this experimentation phase, more sustainable programmes can be developed. Programmes are ongoing and encompass the full life-cycle of digital resources, from selection and creation to management, access and preservation. The rolling of projects into programmes is part of 'the maturing digital library', as described in *The digital library: a biography* [GRE02].

2.1.2 Document classification

As there is a wide range of forms and functions of documents it is relevant to set up a document classification. A classification brings like things together and in this section photographs are placed in a context in order to define their class as clearly as possible. A classification can be based on several principles. Examples are classifications based on the senses used to experience them, classification by medium or classification by subject ([SVE00] pp. 111–113). In several definitions of the

June 2004]. The website of 'Geheugen van Nederland' can be found at: <http://www.geheugenvannederland.nl> [cited 15 June 2004].

[11] See <http://www.rlg.org/visguides> [cited 15 June 2004].

concept ‘document’, the storage medium is an important part. Traditionally documents, defined as recorded information on a stable storage medium, are classified according to technological means of creation and transmission. It is obvious that the characteristics of the analogue storage medium, such as the physical dimension, very much determine the capture and rendition of the graphical material.

The importance of a document classification is illustrated by the conversion problems that may occur if the document is not assessed in the correct manner. It can be difficult, for instance, to make a distinction between a halftone illustration, such as a full-colour illustration in a magazine, and a ‘real’ photograph. The first type of illustration is a photomechanical print. During the digitisation of a photomechanical print the raster of the source can interfere with the raster of the capture device, resulting in the so-called disturbing Moiré effect. A magnifying glass revealing the halftone pattern can be used to assess the document in the correct way. Under normal viewing circumstances the human eye cannot discriminate between a photograph and a modern photomechanical print.

The document classification as presented by Kenney and Rieger ([KEN00] pp. 34–36) is based on the physical characteristics of documents and the technical processes required to create the documents, and consists of the following five categories:

- *Text / line art*. A document with two distinct contrasting colours (often black and white) with no intermediate shades of grey. Printed texts and drawings are part of this category.
- *Manuscripts*. Documents with no clear distinction between foreground and background. The difference between ink and paper, for instance, is gradual. Examples of documents in this category are: hand-written texts and hand-drawn illustrations and charcoal drawings.
- *Halftone or halftone-like items*. A fine-grained pattern of dots makes up a graphical illustration. The pattern of dots gives the human brain the illusion of greyscales or even colour. They are difficult to capture digitally, as the screen of the halftone and the grid of the digital image can conflict, resulting in images with distortions.
- *Continuous tone or continuous-tone-like material*. These documents contain a gradual tone scale from light to dark. The individual tone scales cannot be distinguished. Photographs are part of this category.
- *Mixed material*. These documents comprise more than one of the categories mentioned, such as illustrated pages from a newspaper or book.

2.1.3 Focus on photographs

The word ‘photography’ is derived from the Greek words photos (light) and graphein (writing). Two techniques for ‘writing with light’, introduced in 1839, are considered as the birth of photography. In France L.J.M. Daguerre introduced his daguerreotype, which makes it possible to create unique, very detailed images on a smooth polished plate. In England William Fox Talbot invented a procedure that makes it possible to create a number of photographic prints based on an original negative. In the course of time these techniques were improved and modified. New photographic techniques were invented. Some had more success than others. A number of photographic processes that came into existence are ‘albumen prints’, ‘ambrotype’, ‘salted paper prints’, ‘blueprints’ and ‘wet collodion negatives’.[12] For an overview of the development of photographic techniques, *A guide to early photographic processes* [COE83] and *Care and identification of 19th century photographic prints* [REI86] are useful.

The description and determination of photographic techniques require expert knowledge. On some very specific issues, experts disagree on the suitable determination of a photographic process. Despite the existence of a wide range of photographic processes and applications it is important to define a photograph as clearly as possible. As the documentation of resources is a significant aspect for their longevity, it is important to strive for an unambiguous determination of photographic items. Svenonius supports this observation by stating that ‘What is difficult to identify is difficult to describe and therefore difficult to organize’ ([SVE00] p. 13). Photographs can be considered as documents containing an image that has a continuous tone scale. The following pragmatic classification makes it possible to almost unambiguously identify a photograph, even by non-experts. The classification consists of four categories:[13]

- *Type*: reflective or transparent.
  A photograph is either a medium that reflects light or a medium that allows light to pass through. The first type is called a reflective photograph and the second type is a transparent photograph.
- *Polarity*: positive or negative.
  A photograph is either negative or positive. Negatives contain tonal values opposite from reality. The daylight sky, for instance, on a negative has the darkest tonal values. Photographs with a positive polarity have tonal values that are similar to reality.
- *Colour*: monochromatic or polychromatic.
  A photograph is either monochromatic or polychromatic. Monochromatic means ‘only one colour’. Polychromatic means ‘more than one colour’. So both ‘black and white photographs’ and ‘blue prints’ are monochromatic photographs. A colour photograph is polychromatic.
- *Carrier*: paper, glass, plastic, metal or other.
  The base or primary support of a photograph is either paper, glass, plastic, metal or some ‘other’ medium.

[12] The Thesaurus for Graphic Materials II: Genre & Physical Characteristic Terms contains a large number of photographic techniques. See: <http://lcweb.loc.gov/rr/print/tgm2/> [cited 15 June 2004].

[13] This classification can be found in [HOR01B] and is also part of the SEPIADES data element set (see <http://www.knaw.nl/ecpa/sepia.html> [cited 15 June 2004].

Table 2.1 Three steps in the photographic process

	Conventional photography	Digital photography
Step 1. Capture	'Snapshot'	Scanning / Digital capture
Step 2. Process	Chemical film processing	Digital image processing
Step 3. Output	Imprint	Screen view / Printer output

The identification of a photograph is even more explicit when, in addition to the four aspects mentioned above, another two categories are involved in the classification. These two additional elements are:

– *Dimension*: the length and width of the photograph in a suitable format.
– *Date*: the date the photograph was created. Determining the date when a photograph was created is either very easy, because the date is documented somewhere, or very difficult, because there is no date indication available.

A transparent polychromatic photograph with a positive polarity on a plastic carrier and a width of 35 mm is a more accurate formal description of a specific type of photograph commonly known as a 'slide'. Non-experts can classify other, less commonly known photographic types in the same way. The photograph as a physical object can be very effectively identified by the classification given above.

The photographic process can be divided into three steps (see Table 2.1). First, the image is captured, then the image is processed, and the third step is the creation of an output image. It can be observed that this three-step approach is common to both conventional photography and digital photography. The respective steps are listed in the table. Conventional photography involves the formation of images through the storing of a latent image on a film with the use of light. The film contains sensitive silver compounds that make it possible to make the captured scene visible. In digital photography basically these sensitive compounds are replaced by a charge coupled device (CCD) and the image is not stored on a film but in computer memory.

No exact figures are available on the quantity of photographic items kept by memory institutes. A survey carried out in 1999 among memory institutes in 29 different European countries revealed that they hold an enormous quantity of photographs. The 141 respondents in the survey together keep about 120 million photographic items. The average size of a collection is about 800,000 items. Around half of the photographs in the collections are more than 50 years old ([KLI00] pp. 7-9). Also, there are no detailed figures available on the number of digitised photographs. Mattison estimates that in 2002 about 9 million digital historical photographs are available online [MAT00]. Here, a historical photograph is considered as 20-25 years old. The indicators given above reveal that digitisation of historical photographs is carried out on a wide scale, resulting in millions of digital objects.

2.1.4 Digitisation of historical photographs

This section describes how photographic items can be converted into digital surrogates. As stated in the first section of this chapter, the main reason for applying digitisation techniques is to enhance access to the items and to preserve the originals. The first part of this section deals with the rendering intent of the digital surrogate and the level of preservation of features of the original photograph in the digital surrogate. The second part covers the digital capture process. The functioning of capture devices is described as well as the importance of objective quality parameters.

Rendering intent

Rather than the technical attributes of a digital capture device and the characteristics of the IT infrastructure, it is better to consider the features of the photographs and the rendering intents as the starting point for a digitisation process. Too often the technical infrastructure in terms of storage capacity, data transfer bandwidth, resolving power of the capture device, etc. determines the way sources are digitised. The intended use of the digital surrogate should drive the decisions regarding the digital image quality. Regarding the digitisation of historical photographs, four principles or rendering intents can be distinguished (see: [EST96], [FRE97C] pp. 113-114 and [FRE99] pp. 28-29):

- *The photographic image is rendered*. The digital surrogate should match the appearance of the original photograph 'as is'.
- *The photographer's intent is rendered*. The scanner operator has to adjust the tone and colour reproduction if a photograph is not exposed or processed according to the intention of the photographer. This rendering intent is in principle only relevant for fine-art photography.
- *The original appearance of the photograph is rendered*. If a photograph is faded or has scratches, special image-processing techniques are required to reveal the original appearance.
- *The original scene is rendered*. A 'real life object' is visible on the photograph and the colour of the original must match the colour on the photograph. For this, a calibrated target must be included on the photograph and film and lighting conditions must be benchmarked.

Most digital historical documentary photograph collections are converted according to the first rendering intention.

Levels of preservation

Decisions have to be made regarding the function of the digital surrogate. The two extremes here are on the one hand a visual reference to the original and on the other hand a digital surrogate that can act as a replacement of the original in terms of spatial and tonal information content. This second principle is the closest to the notion of high-quality, use-neutral 'permanent pixels' that facilitate the passive preservation of the vulnerable original. In order to determine to what extent the digital surrogate resembles the original, three levels of preservation can be distinguished. This method is introduced and applied in the publication *Illustrated book study* [KEN99]. The three levels of preservation are:

- *Preservation of structure*. Representing the process or technique used to create the original. The level required for a positive identification of the photographic type varies with the process used to create it. It is easy, for instance, to make a positive identification of a daguerreotype with the unaided eye. The structure of a collotype, however, may only be observable at magnification rates above 25x.
- *Preservation of detail*. Representing the smallest significant part typically observable close up or under slight magnification, again a psycho-visual determination.
- *Preservation of essence*. Representing what the unaided eye can detect at a normal reading distance. This view is based on the psycho-visual experience of the reader rather than any feature associated with the source document.

The *Illustrated book study* concluded that only very high-resolution digital images with a truthful representation of the details and colour range of the original could provide good evidence of structure. There are no guidelines or accepted standards for the creation of digital images according to one of the levels of preservation. In most cases available project resources, such as budget, time, skills and equipment, will dictate the quality level of the digital surrogate.

Image capture devices

With the help of an image capture device a photograph can be digitised. These devices contain a light sensor, a charge coupled device or CCD, which is able to convert the colour and intensity of light to signals represented by binary digits that can be processed by computers. The sensors capture the light that is projected on to a reflective photograph or register light that passes through a transparent photograph. Dark parts on the originals absorb light, as a result of which the sensor does not measure light. The computer code that belongs to this state is '0' (off). Light surfaces of the original reflect light and this is converted by the sensor to the code '1' (on). By extending this principle into two dimensions the working of a digital capture device can be explained.

First of all, in order to register the full tonal scale of a photograph a digital capture device must be able to capture more colours than 'black' and 'white'. Secondly, a digital capture device should have enough resolving power to capture all significant details visible on the original. The number of registered colours and grey levels is expressed as 'bit depth' or 'dynamic range'. It is common to use a minimum of 8 bits to store the colour value of a photograph, enabling 256 (2^8) grey levels. For colour coding a red, green and blue filter is used between the original and the sensor. For each colour, again at least 8 bits are required to have enough unique values, resulting in a bit depth of 24, enabling the coding of 16,777,216 unique colours (2^{24}). The degree of optical opacity, or density, of the original determines the bit depth required to capture its full dynamic range.

The number of times per surface unit the intensity of light is sampled by the digital capture device determines its resolution or resolving power. Three types of device can be used to digitise photographs. For the digitisation of reflective photographs, flatbed scanners can be used. Film scanners are best suited to converting transparent photographs. Both flatbed and film scanners have a fixed sensor dimension, which means that the maximum resolving power of these devices, expressed as the maximum dpi, is fixed. Digital cameras can digitise both reflective and transparent photographs, depending on the light source used. The maximum resolving power of a digital camera is determined by the digital raster size of the camera and the distance between the camera and the original.

The quality of the digital image produced by a capture device is determined by the characteristics of the signal registered by the device. Ideally, the features of the signal should meet the assessed rendering intent and the assessed level of preservation. As a capture device consists of a number of hardware components, such as a light source and optics, a benchmark phase is required to assess the performance of the capture device. With the help of calibration tools the quality of the signal can be measured, such as the colour reproduction, the tone reproduction, the reproduction of the details and the signal–noise ratio. Noise is unwanted variations in the response of an imaging system.

Section 2.1.2 of this study indicates the wide range of photographic formats. Even within a specific classification of photographs, for instance a collection of black-and-white reflective prints, features such as details visible on the image and the contrast will differ. Rather than basing the required quality settings of a digitisation process on the evaluation of the digitisation of an arbitrarily chosen original it is better to use a calibrated greyscale test pattern. This is a photograph consisting of a number of different grey patches for which the optical densities and dynamic ranges are available. It is not easy to carry out this benchmark process, as is illustrated by the quote by Frey in 1997: 'There are no guidelines or accepted standards for determining the level of image quality required in the creation of digital image databases for access or preservation of photographic collections' (FRE97A) p. 597).

In the course of time Technical Committee 42 of ISO, 'Photography', developed a number of standards relevant for the benchmarking of digital capture devices. The scope of ISO/TC42 is standardisation of still picture imaging, both chemical and electronic.[14] ISO/TC42 is developing a number of standards and specifications relevant for measuring performance characteristics of devices used in electronic still imaging. Also, methods, measurements, specifications and recommended

[14] The website of ISO/TC42 – *Photography* can be found at: <http://www.i3a.org/iso.html> [cited 12 June 2004].

practices for storage, permanence, integrity and security of imaging media and materials have been developed by ISO/TC42. The standards created by ISO/TC42 are discussed in section 3.1.

The book *Desktop scanners. Image quality evaluation* [GAN99] is one of the first references for performing quality evaluations of flatbed scanners. A thorough description of digital capture devices as well as quality evaluation features of capture devices can be found in [WIL00].

The Modulation Transfer Function (MTF) of an imaging device is a much better quality indicator than the widely used resolution measurement expressed in 'dots per inch' (dpi). After all, by using interpolation techniques the resolution expressed as dpi can be enhanced artificially. MTF measures the optical frequency response of an imaging system when scanning line pairs whose frequency is within the limits of the scanner. It reveals the optical resolution of an imaging system and how well its optical system performs for a given frequency. It is stated that off-the-shelf software is required to implement a user-friendly quality assessment procedure in a digital conversion workflow. This software must be able to calculate the MTF of a capture device as the best overall measure of detail and resolution. In [WIL98] the meaning and relevance of the MTF are described.

2.1.5 Digital images

According to Kirsch [KIR98] the computer processing of images in the USA began at the National Bureau of Standards in 1956. Several decisions made by the developers of the first scanner have influenced engineering practice ever since, for instance the usage of rectangular arrays of square pixels. No attempt was made to base the digitisation protocol on the nature of the image, leading to rather large images. Kirsch illustrates this fact by showing a sixth-century 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, for instance, text, spreadsheet or numerical data) and that are intended for eventual rendering and display ([MUR94] p. 5). The *Encyclopaedia of graphics file formats* [MUR94], published in 1994, describes almost 100 different file formats. 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 fundamentally different types of graphical data. Each type requires its own file format. Browne and Sheperd ([BRO95] p. 4) make a distinction between three broad categories:

- *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.[15]

[15] Raster data is also called Bitmapped data.

- *Geometry data*: mathematical description of space, in either 2-dimensional or 3-dimensional space, that represents the components of an image.[16]
- *Latent image data*: non-graphical data that can be transformed into useful images by some algorithmic process.

In this study the digital surrogate of an historical photograph is in all cases 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 ([MUR94] pp. 384-387) is an example of a format that was originally proprietary, but at a later stage Kodak (more or less forced by the market, which was demanding more open formats) permitted developers to use the format specifications. Some graphics file formats were directed towards 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 of processing the image and the memory required to represent the image. The books *Graphics file formats* [BRO95] and *The file formats Handbook* [BOR95] are a rich source of information on the specification of image file formats. The goal of the books is to help to understand how graphic data streams differ and why.

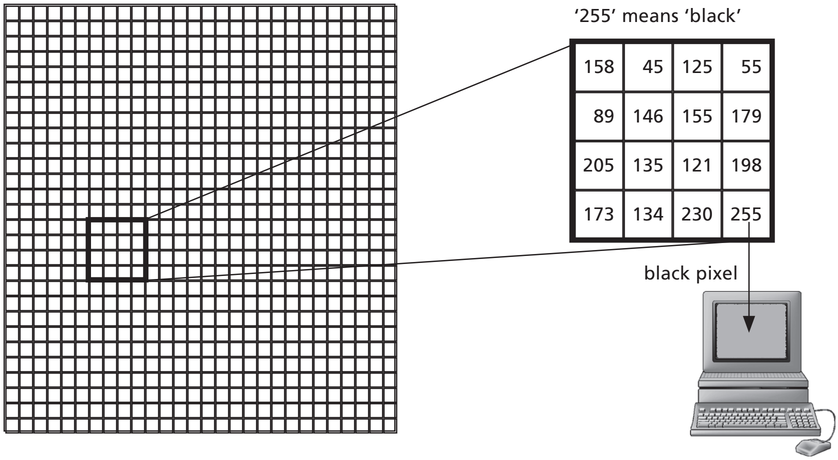
The design of a graphics file format is based on the memory, speed and circuitry components of the hardware systems targeted to use the data. The amount of available memory can affect the speed of data access. Graphics applications are real memory hogs. Device independence decreases processing speed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

Raster image files

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 is a light-sensitive sensor on a chip for converting the analogue signal into discrete digital codes. The sampling interval of the image capture device determines how many pixels are stored in the bitmap file. These pixel codes can be written in a number of ways. The bitmap files can contain additional data such as image description information or a colour palette.

The memory capacity in bits required to store the pixels of a bitmap can be expressed by the formula: PH X PV X PD. 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 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 ([FRE99] p. 21). The comparison between

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



158	45	125	55
89	146	155	179
205	135	121	198
173	134	230	255

Figure 2.1 Raster image file, a matrix of discrete pixel values

silver halide and CCD sensors given in ([JAC00] p. 21) confirms that photographic materials have a very high resolving power. Depending on the rendering intent and required level of preservation the representation of a photograph as a bitmap can result in huge raster files.

Image processing

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 (see Figure 2.1). 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 (see [UMB98] pp. 5-7).

Data compression algorithms are applied in order to reduce the file size (see [MUR94] pp. 125-171). As bitmap files tend to be very large, 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 no longer have the discrete raster structure and the compression method usually results in a loss of visual 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 image files in the TIFF data format. Regarding the durability of bitmap files, image-processing algorithms can be subdivided into two groups:

- *Image processing algorithms that change values of the pixels in the bitstream of the stored bitmap*. Examples are: image compression and modification of the pixel values by digital filters and contrast and brightness adaptations. In addition, image compression often abandons the rastered structure of the bitmap, as illustrated in Figure 2.1.

– *Image processing algorithms that do not change the bitstream of 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 the RAM memory of the computer and do not permanently manipulate the pixel values in the bitmap.

Image-processing algorithms that change the pixel values in the bitmap are best not used, because the characteristics of the base bitmap are changed. Future image processing for specific purposes cannot be founded on a 'use-neutral' master file. For a number of reasons the use of compression algorithms is not recommended. There is a risk of the compression method becoming obsolete and, if the bitmap becomes corrupted, it is very difficult to repair the image. An uncompressed image can be repaired much more easily. Lastly, the application of image compression usually leads to a loss of image quality.

Image file formats

Browne and Sheperd [BRO95] evaluate a number of design goals for graphic file formats that determine the type of data format that is supported, the types of data encoding that will be used and the overall data organisation of the format. These design goals determine the way in which a bitmap file is created, read and written. Longevity or durability is not mentioned as a specific design goal of a bitmap format. In principle, it is possible to design a new, specific 'permanent pixels' graphic file format whose main goal is to guarantee long-term access and use. Fast rendering and efficient memory usage will obviously not be as important as the robust encoding (for instance, by using the ASCII standard) of the pixels that make up the image. A new specific format should only be designed if none of the current existing file formats can be considered as durable.

Based on Browne and Sheperd, two data formats seem to fit as a building block to create 'permanent pixels', namely TIFF (tagged image file format) and CGM (computer graphics metafile). Since the authoritative book of Browne and Sheperd was published in 1995 the TIFF standard has been used much more widely than CGM. Despite the fact that CGM is directed towards the support of both raster and geometric data it seems that current implementations support mainly the geometric data set. TIFF is a standard developed by industry and has a huge user community. CGM is an official ISO standard developed by ISO/IEC JTC1 subcommittee 24 for 'encoding picture data' [ISO8632:1994].

The TIFF image file standard seems to be the most appropriate candidate for use as a component for 'permanent pixels'. Frey [FRE00B] states, 'From the currently available formats, TIFF is the one that can be considered most 'archival'. It is a versatile, platform-independent and open file format and it is being used in most digitising projects as the format of choice for the digital masters'. TIFF is a tag-based file format designed to promote universal interchanges of digital image data. As TIFF files do not have one specific way to store image data, there are many versions of TIFF. Therefore the TIFF standard must be evaluated more closely. This is done in section 4.1 of this study.

2.1.6 Digital master files

The way in which an organisation decides to digitise its photographic collection is determined by a number of factors. The characteristics of the collection, the budget and skills available and the rendering intent and level of preservation are the main parameters that determine the digitisation approach. *Moving theory into practice, digital imaging for libraries and archives* states: 'although there is no consensus on the appropriate method for determining requirements, there is growing support for creating digital masters that are rich enough to be useful over time and cost-effective' ([KEN00] p. 24). The quality of the initial scan must be as high as possible. Benchmarking the conversion chain leads to informed decisions on a range of issues. Frey states: 'The creation of large digital image collections is not likely to be attempted more than once a generation. This means that it had better be done right the first time, so being aware of the technical nature of the digital images produced is quite important' ([FRE97A] p. 599). Within memory institutes a number of conversion requirements are developed. Files that are considered as digital masters can have different properties.[17] This means that an objective image quality assurance for digital surrogates of historical photographs does not exist. In the electronic publishing world the situation is much better. The 'graphic-design-to-printing-press' chain contains a number of methods for achieving objective, predictable quality. These methods are described in [TAL99].

A number of factors determine the quality of a digital image, such as the characteristics of the original to be digitised, the rendering intent, the required preservation level and the features of the image capture device. Ultimately a human being will see the digital image rendered on a computer screen or printed on paper. These viewing conditions can vary a lot and thus influence the quality perception of the viewer. Standard ISO 3664 *Viewing conditions – Graphic technology and photography* [ISO3664:2000] provides specifications for illumination and viewing conditions for images on both reflective and transmissive media. This normative reference can be used for appraisal of images on monitors and on hard copy, such as prints and transparencies. The standard contains guidelines for judging and exhibiting photographs, such as requirements for colour temperature and illumination levels.

Figure 2.2 contains a model of the aspects that play a role in the creation of durable digital surrogates of analogue historical photographs. Based on a classifica-

[17] A table with conversion requirements from a number of institutions can be found at: <http://www.library.cornell.edu/preservation/tutorial/conversion/table3-1.html> [cited 15 June 2004]. This table is part of the online 'Digital imaging tutorial' that is related to [KEN00].

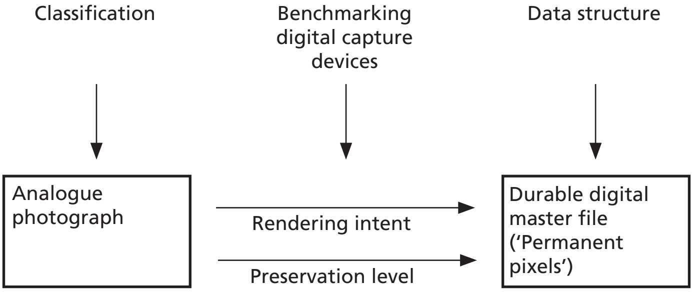

Classification
Benchmarking digital capture devices
Data structure
Analogue photograph
Rendering intent
Preservation level
Durable digital master file ('Permanent pixels')

Figure 2.2 Durable digital surrogates of analogue historical photographs

tion, the physical characteristics of the analogue source material can be assessed. This classification is required in order to determine which digital capture device is the most appropriate to use. The settings of the actual digitisation process are based on the assessed rendering intent and required preservation level. The digitisation is based on a benchmarking phase in which digital capture devices are evaluated. The result of the digitisation process is stored in a specific data structure – a raster image file. The data organisation and pixel encoding mechanism of this raster image file is an important indicator of the durability of the digital master image. Future operating systems, data storage formats and applications are among the most important factors that threaten long-term use of the digital master file.

The translation of the rendering intent and assessed preservation level into digital capture requirements is obstructed by two factors. First, the tone scale and level of detail fluctuate within a collection of photographs, even if all photographs have the same classification. This means that in principle, for each individual analogue original, dedicated digitisation settings have to be determined and this is not feasible in practice. This is the reason why functional, practically compromised digitisation settings are formulated. These settings differ between institutes and they can at best be considered as directives. The three main aspects of these settings are the pixel depth, the pixel density and the image file format. The pixel depth setting must enable the capture of the dynamic range of the analogue original and the pixel density must enable the capture of the details of the analogue original.

The second reason for the difficulty in arriving at objective digital conversion settings is that the standards and tools for characterising the performance of digital capture devices require the specialised skills of an imaging scientist. This situation is changing, as is illustrated by the telling keynote address at a scientific imaging conference with the title *Sneaking scientific validity into imaging tools for the masses* [BER02]. The keynote address covers the question of how typical imaging practices can become scientific imaging practices. If this is the case, the performance of the digital capture device is based on objective standards and tools. These tools and standards are starting to emerge and require attention from both the imaging scientists and the professionals in the memory institutes that create digital durable masters. One of the use cases described by Berns illustrates what should eventually be possible: 'An archivist is expected to digitise photographic reproductions such that the digital archive is an accurate colour reproduction of the original work of art' ([BER02] p. 1).

2.2 Digital longevity in memory institutes

Increasingly, the preservation management of digital materials is becoming a structural activity within memory institutes. The number of digital artefacts is growing and long-term access and maintenance of digital data are gaining importance. Digital surrogates of photographic items are part of the larger digital preservation framework that is described in more detail in this section. The goal of this section is to present a general preservation framework that places the preservation of a specific digital object, such as the digital surrogate of an historical photograph, in a broader perspective.

This section consists of six parts. First, as an introduction, a general overview is given of digital preservation initiatives in memory institutes. Also part of this introduction is a description of the way the National Archives in the USA are trying to prevent the loss of digital images. The second part of this section covers a reference model for an Open Archival Information System (OAIS) [ISO14721:2003], which is being used increasingly by a lot of memory institutes as a basis for activities in the field of digital preservation. The model contains requirements for an archive to provide permanent, or indefinite long-term, preservation of digital information.

The third part of this chapter section elaborates on preservation metadata. Preservation metadata is documentation that supports the durability of digital artefacts of memory institutes. Often the elements that are part of a metadata schema are 'mixed and matched', resulting in so-called 'application profiles'. This issue is covered in part four of this section. The fifth part of the section covers the relevance of the eXtensible Markup Language (XML) standard as a future-proof, non-proprietary data storage format. The sixth part of the section discusses business models for digital preservation.

2.2.1 Digitisation projects and programmes of memory institutes

The Dutch poet Lucebert (1924-1994) wrote, 'everything of value is defenceless'.[18] An attempt to interpret this quote might result in the observation that special attention is required to enable the future survival of 'valuable things'. The holdings of archives, libraries and museums contain a huge number of valuable things and the conservation of these objects is an important task of the institutes. More and more memory institutes are creating, collecting and receiving digital objects and the permanency of these objects is increasingly considered as an important issue.

Whereas analogue assets such as paintings or furniture gradually deteriorate

[18] English translation of the original Dutch poem line 'alles van waarde is weerloos'.

in the course of time, a digital object is threatened by sudden damage or loss, because the bitstream that makes up the digital object can no longer be processed. This 'digital cliff' requires active involvement by memory institutes to assess their collection of digital assets and actively carry out a digital preservation policy for objects that are considered valuable enough for future generations. As digitisation projects require considerable investment, it is apparent that long-term usage of the digital objects is required.

An important source of information on the creation and management of digital materials over time is the handbook *Preservation management of digital materials* [JON01]. Via a website, developments in the field of digital preservation are available that could not be printed in the book.[19] The handbook reports on the current thinking on digital preservation issues and gives attention to the various stages in the life-cycle of digital materials. Within the cultural heritage community a number of organisations are carrying out a wide range of activities relevant for the preservation of digital objects. These activities range from raising awareness to developing standards and coordinating projects. Examples of such organisations are given below. Each member of the 'memory community' is represented by an international and a Dutch example.

– *Libraries*. Within the library community the Research Libraries Group (RLG) carries out activities under the caption 'Long term retention of digital research materials'.[20] In the Netherlands the Royal Library in The Hague has developed an electronic deposit system for preserving electronic publications. See [STE02] and [DIE02].[21]
– *Archives*. The European ERPANET project,[22] whose partners originate mainly in the archival world, is aiming to establish an expandable and self-sustaining European Initiative, which will serve as a virtual clearinghouse and knowledge base in the area of preservation of cultural heritage and scientific digital objects. In the Netherlands the project 'digitale duurzaamheid' (digital longevity), initiated by the Dutch National Archives and the Ministry of the Interior, has carried out projects to test alternatives for the digital archiving of government documents.[23]

  Scientific digital data archives are responsible for long-term access to repositories of datasets created by scholars. One of the first scientific data archives is ICPSR (Inter-university Consortium for Social and Political Research),

[19] The website can be found at: <http://www.dpconline.org/graphics/whatsnew/> [cited 15 June 2004].

[20] The website of this project is: <http://www.rlg.org/longterm/index.html> [cited 15 June 2004].

[21] The website of the KB/IBM long-term preservation study can be found at: <http://www.kb.nl/hrd/dd/dd_onderzoek/dnep_ltp_study-en.html> [cited 12 September 2004].

[22] The website of ERPANET (Electronic Resource Preservation and Access NETwork) can be found at: <http://www.erpanet.org> [cited 15 June 2004].

[23] The website can be found at: <http://www.digitaleduurzaamheid.nl> [cited 15 June 2004].

founded in 1962. From 1989 onwards the NHDA (Netherlands the Historical Data Archive) is active as part of a number of umbrella organisations [DOO89], [DOO90] and [DOO96].

– *Museums*. As the main task of museums is to exhibit their holdings to the public, organisations focusing mainly on the preservation of digital museum materials are hard to find. The Canadian Heritage Information Network (CHIN) maintains a best practice guide on digital preservation for Museums.[24] From 1990 until 2003 the Consortium for the exchange of museum information (CIMI) existed, which worked on a standard for the description of the information categories that can be used when creating electronic records about the objects in museum collections [GUI00b]. Digital preservation was an issue covered by this initiative. A Dutch initiative that provides electronic access to digital surrogates of cultural heritage objects is 'Geheugen van Nederland' (Memory of the Netherlands).[25] Museums are among the main suppliers of content for this digital collection.

These examples illustrate that the electronic access of digital objects is on the agenda of memory institutes and that the longevity of these objects is important mainly for the archive and library community. A broad initiative relevant for all memory institutes is the research cluster 'Digital Preservation' of the DELOS network of excellence on digital libraries, funded by the EU Sixth Framework Programme.[26] This research cluster aims to coordinate and promote research on digital preservation on a European scale.

Archiving of 'photographic records' at NARA

The National Archives and Records Management (NARA) in the USA is one of the first organisations that accepted the responsibility for long-term access to digital surrogates of historical photographs. As an example, the approach of NARA is described in more detail. The practices related to the archiving of digital images by NARA can be used as a reference for further detailed research, as presented later in this study.

Under the framework of the Electronic Records Management Initiative, NARA has developed tools that agencies will need for managing their records in electronic form. The project provides guidance on electronic records management and will enable agencies to transfer electronic records to NARA in a variety of data types and formats so that they may be preserved for future use by the government and citizens. In 2003 NARA expanded the number of formats as well as the media and techniques that can be used by agencies. One of the permanent electronic records

[24] The 'Best practice for Museums. Digital preservation' can be found at: <http://www.chin.gc.ca/English/Digital_Content/Digital_Preservation/> [cited 15 June 2004].

[25] The website can be found at: <http://www.geheugenvannederland.nl> [cited 15 June 2004].

[26] The website of the digital preservation cluster of the DELOS project can be found at: <http://www.delos.info/wp6.html> [cited 25 September 2004].

supported by NARA are digital photographic records. A transfer instruction has been compiled that contains the requirements for transferring this type of permanent electronic record to NARA.[27]

The guidance applies to digital photographic records that have been appraised and scheduled for permanent retention at NARA. Included under the scope of this guidance are still photographs of natural, real-world scenes or subjects created in support of agency business that are produced by digital cameras, as well as scanned images of photographic prints, slides and negatives. The guidance applies to master image files of digital photographs created using medium- to high-quality resolution settings appropriate for continued preservation. Low-resolution images, images captured with office automation applications, aerial photography and satellite imagery, and vector-based images are not accepted for permanent storage at NARA.

As of November 2003 the permanent storage of digital photographic records has been operational at NARA. Additional requirements have applied to permanent digital photographic records from January 2005 onwards. NARA will accept digital photographic records in the following file formats and versions. Additional formats may be added in the future.

– *Tagged Image File Format* (TIFF), versions 4.0 (April 1987), 5.0 (October 1988) and 6.0 (June 1992), published as [TIF92]. Default file name extensions include .TIF and .TIFF. NARA requires TIFF formatted images in which the byte order is always from the least significant byte.
– *JPEG File Interchange Format* (JFIF, JPEG), all versions compliant with [ISO10918-1:1994]. Default file extensions include .JPEG, .JFIF and .JPG.

Images must be provided as continuous-tone greyscale or colour raster images, 8-bit or 16-bit per channel. Colour images must be produced in RGB (Red Green Blue) colour mode as 24-bit or 48-bit colour files. Records created using digital cameras must be captured as 2 megapixel files or greater with a minimum pixel array of 1,600 pixels by 1,200 pixels. Since 1 January 2005, digital camera files must be captured as 6 megapixel files or greater with a minimum pixel array of 3,000 pixels by 2,000 pixels. Records produced at this resolution and size are comparable in quality to 35-mm film photographs, which is the minimum quality level for still pictures currently accepted by NARA.

Since 1 January 2005, agencies must ensure that digital cameras and scanners produce records with true optical resolution. Resizing images or interpolating to a higher resolution from a lower resolution for purposes of transfer will not be permitted. NARA will accept digital photographs in TIFF file formats that are compressed using a loss-less compression method. If available, NARA prefers that agencies transfer uncompressed versions of these files. NARA will accept permanent digital photographs in the JPEG file format, which uses a lossy compression

[27] This guidance is available online at: < http://www.archives.gov/records-mgmt/initiatives/digital-photo-records.html > [cited 8 January 2004].

method, provided that the records have been created using at least medium-quality compression settings.

Agencies must transfer to NARA first-generation JPEG files that have not been degraded in quality by multiple revisions and re-saving. Making changes to JPEG files (for instance, altering the image size), and then re-saving them, can result in re-compression of the images, leading to additional data loss and degradation of image quality. Agencies must provide descriptive information about the records. NARA prefers that this information be captured for each image in the image header, but will accept transfers of records with this information in the accompanying documentation. Specific information required includes, but is not limited to:

- *Unique photograph identification number.* Identify each individual photograph with a unique identification number and/or file name. If agency-specific naming conventions are used, documentation must be provided describing these standards.
- *Caption.* Provide narrative text describing each individual image in order to understand and retrieve it. Standard caption information typically includes the 'who, what, when, where, why' about the photograph.
- *Photographer.* Identify the full name (and rank, if military) and organisation of the photographer credited with the photograph, if available.
- *Copyright.* Indicate for each image whether there is a restriction on the use of that image because of copyright or other property rights. Agencies must provide, if applicable, the owner of the copyright and any conditions on the use of the photograph(s), such as start and end dates of the restriction.

Agencies must also provide technical information about the records. NARA prefers that this information be captured for each image in the image header, but will accept transfers of records with this information in the accompanying documentation. Specific information required includes, but is not limited to:

- *File format.* List the file format and version of each image file transferred to NARA.
- *Bit depth.* Identify the bit depth of the transferred files.
- *Image size.* Specify the image height and width of each image in pixels.
- *Image source.* Identify the original medium used to capture the images (for instance, the make and model of the digital camera or the make and type of the film used).
- *Compression.* Identify the file compression method used (if applicable) and the compression level (for instance, medium, high) selected for the image(s).
- *ICC/ICM profile.* Provide custom or generic colour profiles, if available, for the digital camera or scanner used (for instance, sRGB: Standard Red Green Blue).[28]
- *EXIF information.* If available, preserve and transfer to NARA the Exchangeable Image File Format (EXIF, see [EXI02]) information embedded in the

[28] More information on the colour coding of digital images can be found on page 91 *et seq.*

header of image files (as TIFF tags or JPEG markers) by certain digital cameras, for instance make and model of the digital camera.[29]

For digitised analogue photographs, agencies must supply a description of the quality control inspection process, a report on the results of the last inspection performed on the records and the date of that inspection. As part of the report, agencies should visually inspect a sample of the images for defects, evaluate the accuracy of finding aid data and verify file header information, as well as file name integrity and completeness of the images in the transfer.

NARA will provide access to the creating agency and to all researchers requesting digital photographic records accessioned from federal agencies. While compliance with these requirements will improve future access to accessioned digital photographic records, NARA's ability to provide access to certain records will vary according to their hardware and software dependencies. For digital photographic records transferred to NARA, the user will be responsible for obtaining the necessary hardware and software to view the records.

Documentation of the analogue source is an essential part of the metadata that supports the longevity of the digital surrogate. The better the original source is represented both in the metadata and the digital surrogate, the less the vulnerable original has to be touched. Thus the digital surrogate of a source is part of the (passive) preservation policy of the original. Despite the fact that NARA actively facilitates the durability of digital images, a number of issues are defined in a global, open-ended way and this justifies further research.

2.2.2 Reference Model for an Open Archival Information System (OAIS)

The *Reference Model for an Open Archival Information System* (OAIS) establishes a common framework of terms and concepts relevant for the long-term archiving of digital data. The OAIS reference model has been developed under the direction of the 'Consultative Committee for Space Data Systems' (CCSDS) and adopted as ISO standard 14721 [ISO14721:2003]. 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' ([ISO14721:2003] pp. 1-11). 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' (ISO14721:2003] pp. 1-10). The OAIS model is widely used as a foundation stone for a wide range of digital preservation initiatives. The 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 acts as a de-facto standard that is used on a large scale for digital preservation. The core values of archival activity are the sanctity

[29] More information on the EXIF file format can be found on page 65 *et seq*.

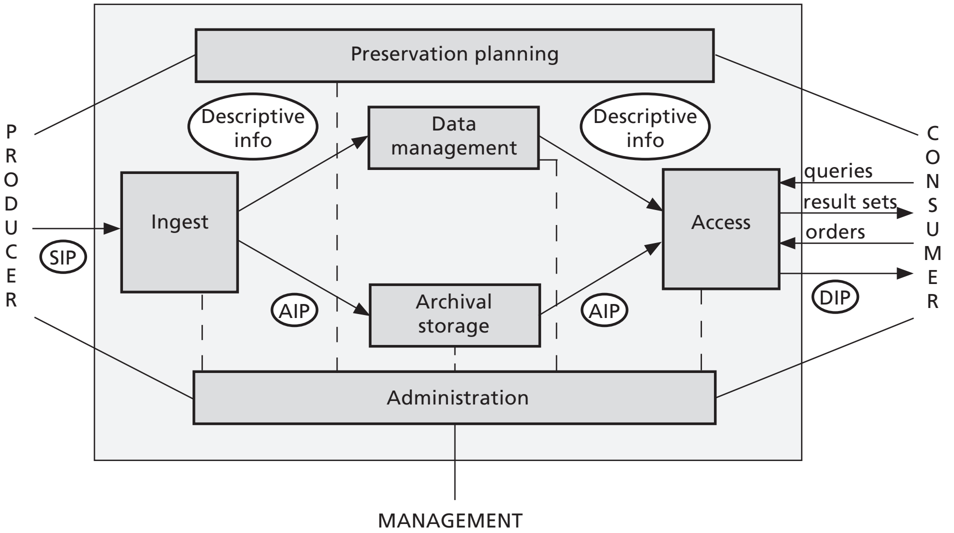


Figure 2.3 OAIS Functional Entities (Source: [ISO14721:2003] p. 4-1)

of evidence, the preservation imperative, the primacy of the record, 'respect des fonds', original order and provenance, and hierarchy in records and their collective description. These traditional archival principles and practices are defined and then translated into digital repository architecture designs through an analysis of the OAIS model.

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*. In order to understand the Data Object (this is either a physical object or a digital object) that has been archived, Representation Information is required. 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 by, an OAIS. These variants are referred to as the 'Archival

Information Package' (AIP), the 'Submission Information Package' (SIP), 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 SIP to the OAIS for use in the construction of one or more AIPs. A Consumer receives a DIP, derived from one or more AIPs, in response to a request to the OAIS.

The OAIS functional model consists of six entities:

1. *Ingest*. Contains the services and functions that accept the SIPs from producers, prepares the AIPs for storage, and ensures that the AIPs 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 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 that make the archival information holdings and related services visible to Consumers.

Figure 2.3 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 schemas. The Metadata Encoding and Transmission Standard (METS),[30] for instance, consists of a number of objects that can be seen as an implementation of the OAIS Information Packages SIP, AIP and DIP. METS is an XML Schema that can be used for the encoding of descriptive, administrative and structural metadata regarding objects within a digital library [SEA02].

Secondly, the OAIS model plays a role in the architecture and design of digital preservation information systems. The OAIS standard 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' ([ISO14721:2003] pp. 1-3). The system requirements of the digital deposit service of the National Library of the Netherlands called e-Depot are, for instance, based on the OAIS model ([WIJ04] p. 254).

Thirdly, the OAIS model is used as a basis for conformance and many digital

[30] The website of the METS standard can be found at: <http://www.loc.gov/mets> [cited 24 August 2004].

preservation information systems claim OAIS compliance. However, a generally accepted OAIS certification process does not yet exist. RLG and 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.[31]

OAIS mandatory responsibilities

The OAIS standard distinguishes six mandatory responsibilities that an organisation must discharge in order to operate an OAIS archive ([ISO14721:2003] p. 3-1). The archive 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.

Producer–Archive interface

The relationship and interactions between an information Producer and an Archive (see Figure 2.3) can be problematic. Four potential problems mentioned by the abstract standard *Producer–Archive interface methodology* [PRO04] are divergent expectations of the digital objects, unclearly defined digital objects, an incompletely fulfilled ingest schedule by the Producer, and transfer errors detected too late by the Archive. In line with the OAIS Reference Model, the abstract standard describes the Producer–Archive interactions in a detailed way.

One of the actions that should be carried out in the preliminary phase according to [PRO04] is to 'identify the Designated Community'. It should be made as explicit as possible how and by whom the digital objects will be used. The determination of the Designated Community is not easy, as illustrated by the quote: '... it should be noted that for some institutional and/or governmental Archives neither the Producer nor the Archive has a precise idea of how the information to be preserved will be used. Even with scientific observation archives, 10 years after production, scientific data is used in ways that the Producer could not even imagine'

[31] The website of the RLG/NARA task force on Digital Repository Certification can be found at: <http://www.rlg.org/longterm/certification.html> [cited 15 September 2004].

([PRO04] pp. 3-6). The next section contains an attempt to identify Designated Communities relevant for the long-term archiving of digital surrogates of historical photographs.

2.2.3 Metadata schemas for digital preservation

Without documentation a digital object is just a sequence of binary digits. The longevity of digital objects will improve if 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 about 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. The *Preservation Management of Digital Materials Workbook* states that 'there are factors which make documentation particularly critical for the continued viability of digital materials and they relate to fundamental differences between traditional and digital resources' ([JON01] p. 133).

Literally, metadata means 'data about data'. The library and information science literature published in recent years acknowledges the significant role of metadata for the management of digital objects. See, for instance, [ALE00], [DAY01], [JON01], [LAZ01], [GON04] and [WEI96]. Metadata can be considered as documentation that provides information on the characteristics of 'things', both analogue and digital. Increased attention to the importance of metadata has been generated by the growth of the Internet. The Internet is not a well-organised and structured library. The objects in the Internet information space can only be discovered if 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.

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource ([UND04] p. 1). Metadata is commonly used for any formal scheme or description applying to any type of object, digital or non-digital. Metadata can describe resources at any level of aggregation, such as collections or individual items. Metadata can be embedded in a digital object or it can be stored separately.

The three main types of metadata are descriptive metadata, structural metadata and administrative metadata. Descriptive metadata is used to identify and discover resources. Structural metadata defines the relation of individual objects that form an entity, such as the chapters of a book. Administrative metadata helps to manage a resource, for instance, to settle the intellectual property rights of a resource. Preservation metadata can be considered as a subset of administrative metadata. Preservation metadata contains information needed to archive and preserve a resource. Specific information is required to track the origin of a digital object, such as its format characteristics.

The Dublin Core Metadata Element Set

For one specific function of metadata, namely resource discovery, the 'Dublin Core Metadata Element Set – DCMES' ([ISO15836:2003]) is used on a large scale.[32] The fifteen core elements of DCMES are applied in a large number of projects and initiatives in order to enable the discovery of objects on the Internet. In 2001 DCMES became an official ANSI/NISO standard (Z39.85) and in 2003 DCMES was issued as international ISO standard 15836.

In the article *A grammar of Dublin Core* [BAK00], the Dublin Core data element set is presented as a language. It is a 'pidgin' language because basic words and simple sentence patterns are used. 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'. [DAY01] reviews initiatives that relate to preservation metadata for digital objects.

An important issue concerning the creation of metadata is the clear definition of the resource that is being described. Confusion may exist between the description of the object and the description of the digital surrogate of that object. Also, the intellectual content of the object can have its specific metadata. The scene visible on an original photograph is the same as the rendered scene of the digital surrogate. Next, confusion can exist between item-level descriptions and collection-level descriptions. These definition or determination issues were already covered in the 1996 article *Image description on the Internet* [WEI96]. An often-applied solution for this problem is the employment of the so-called 1:1 principle. The 1:1 principle states that only one object, resource or instantiation may be described within a single metadata record. Surrogates of resources, too, must be described separately from the original object, such that a metadata record for a photograph of a Greek sculpture should contain metadata about the photograph, not about the sculpture (see [GUI00B]).

OAIS metadata schemas

A number of initiatives are being undertaken to define metadata schemas for digital preservation. The OAIS reference model plays an important role in these initiatives. Among the various international projects concerning preservation metadata the CEDARS project, active from 1998 until 2002, was one of the most important. CEDARS explored issues and practical exemplars for the long-term preservation

[32] The website of the Dublin Core metadata initiative is: <http://www.dublincore.org> [cited 24 September 2004]. The fifteen Dublin Core elements are: creator, title, subject, contributor, date, description, publisher, type, format, coverage, rights, relation, source, language and identifier.

of digital materials.[33] Its distributed digital archive architecture is based on an implementation of the OAIS reference model, using a framework that facilitates different Information Packages as defined in the OAIS reference model. The project resulted in recommendations and guidelines for digital preservation. The criteria for preservation metadata as implemented by the CEDARS architecture ensure that the preserved digital object can be found and ensure that the preserved digital object can be understood.

The goal of the PREMIS project is to develop best practices and recommendations for implementing preservation metadata for digital objects.[34] Many different metadata schemas are being developed in a variety of user environments and disciplines. Some of the most common ones relevant for the documentation of digital surrogates of historical photographs are discussed in the next section. A number of Designated Communities, related to memory institutes, are distinguished.

Designated Communities and metadata

Based on relevant projects and literature, a number of Designated Communities can be identified, as defined in the OAIS reference model, capable of using and understanding specific information related to the durability of both digital and analogue historical photographs (see Table 2.2).

The first Designated Community is the group interested in the historical photograph as a physical object. For archivists the photograph is an archival record. Archival description principles are relevant in this respect. Photo historians are basically interested in the photographic technique just like conservators of photographic materials. The second Designated Community is grouped around the digital surrogate or digital image of the photograph. The users of digital imaging techniques and the digital preservation community are part of this group. A third Designated Community are the consumers of the content or scene visible on the image, irrespective of the information carrier. All types of researchers, for instance historians or social scientists, are part of this community.

In contrast to the first three Designated Communities that have been distinguished, the fourth Designated Community is not constructed around its primary object of interest, but takes the interest in photographic collections as its point of departure. Table 2.2 contains examples of what the given Designated Communities possibly expect from the documentation of the historical photographs and their digital surrogates. The last column of Table 2.2 contains a number of metadata schemas. Further research and coordination with the Designated Communi-

[33] The website of the 'Consortium of University Research Libraries (CURL) Exemplars in Digital Archives' (CEDARS) project can be found at: <http://www.leeds.ac.uk/cedars> [cited 15 August 2004].

[34] PREMIS stands for: 'PREservation Metadata: Implementation Strategies', see: <http://www.oclc.org/research/projects/pmwg/> [cited 15 August 2004]. The project is based on the OAIS reference model.

Table 2.2 Some Designated Communities, metadata requirements and candidate metadata schemas

Designated Community	*Primarily interested in:*	*Metadata should give information on:*	*Metadata schemas that possibly contain relevant data elements*
Archivists	Photograph	Relation of photographic item with archival context	International Standard for Archival Descriptions (ISAD)
Photo-historians		Photographic techniques	Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms (TGM II)
Imaging experts	Digital image	Digital image specifications	Data Dictionary - Technical metadata for digital still images (NISOZ39.87:2002]). Exchangeable Image File Format (EXIF)
Archivists			
Historians	Content	Scene visible on the image	Visual Resources Association Core Categories (VRA Core)
Art historians			Iconclass
Institutes and individuals with photographic collections	Multi-level relation between collection and photographic objects (digital / analogue)		SEPIA Description Element Set (SEPIADES)

ties is required in order to establish a consensus on the way the shared and formal description of the concepts is covered by the metadata elements that are part of a metadata schema. Another issue to take into consideration is a critical review of the necessary description elements required to meet the durability criteria.

The example metadata schemas mentioned in the third column of Table 2.2 are:

– The *International Standard Archival Description* (ISAD)[35] is a standard that provides general guidance for the preparation of archival descriptions. The purpose of archival descriptions is to identify and explain the context and content of archival material in order to promote its accessibility. A description of the application of the standard can be found in [HOR99].
– The *Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms* (TGMII)[36] is a thesaurus of more than 600 terms developed by the Library of Congress Prints and Photographs Division with input from other archival image repositories. New terms are added regularly.
– *Data Dictionary – Technical metadata for digital still images*_

[35] Currently the second edition, published in 1999, of the standard is available. The standard can be accessed on the website of the International Council on Archives: <http://www.ica.org> [cited 13 September 2004].

[36] The TGM II is available on the Internet, see: <http://www.loc.gov/rr/print/tgm2/> [cited 24 August 2004].

([NISOZ39.87:2002]) has two fundamental functions. The first function is to document the history and provenance of a digital image. The second function involves the assurance that a digital image will be rendered accurately on output in the long term. As the technical metadata of a digital object is one of the most important types of preservation metadata, later in this study more attention will be paid to this standard. Also, the *EXIF standard* ([EX102]), used to standardise technical metadata created by digital cameras, is covered in more detail.

- Version 3 of the description elements or 'Core Categories' designed by the Visual Resources Association (VRA), known as *VRA Core Categories*, was published in 2002.[37] The schema is directed at the description of works of visual culture as well as the images that document them. A work is a physical entity that exists, has existed at some time in the past, or that could exist in the future. It might be an artistic creation such as a painting or a sculpture. According to the VRA Core Categories an image is a visual representation of a work. It can exist in photomechanical, photographic and digital formats.
- *Iconclass* is a subject-related international classification system for iconographic research and the documentation of images. Iconclass contains over 28,000 definitions, consisting of an alphanumeric classification code, used to index the iconographic contents of works of art. The definitions are used to describe the subjects of images represented in many works and documents ranging from paintings and manuscript illuminations to photographs, posters and newspaper clippings[38].
- The *SEPIA Data Element Set* (SEPIADES) is a set of more than 400 data elements to describe photographs. SEPIADES was compiled during the SEPIA (Safeguarding European Photographic Images for Access) project that ran from 1999 until 2003.[39], [40]

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

Technical metadata is only a subset of the complete suite of preservation metadata elements 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. First, 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. The draft standard NISO Z39.87 *Technical Metadata for Digital Still Images*

[37] The metadata schema VRA Core Categories (3rd version) is available at: <http://www.vraweb.org/vracore3.htm> [cited 25 August 2004].

[38] The homepage of Iconclass is: <http://www.iconclass.nl> [cited 25 August 2004].

[39] For more information on SEPIADES, see: [KLI03].

[40] Another metadata initiative worth mentioning is 'Voor de zoeker', published in 1994. This is one of the few Dutch manuals dedicated to the documentation of photographic collections [HOG94].

[NISOZ39.87:2002] was initiated by the cultural heritage community and is intended mainly for the formulation of technical metadata of digital surrogates of analogue originals.

The purpose of the [NISOZ39.87:2002] data dictionary is to define a standard of metadata 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 [NISOZ39.87:2002] standard are cultural institutions, publishers, rights holders and other organisations engaged in digitising visual materials from archival collections. The metadata 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 [NISOZ39.87:2002] data dictionary covers four categories of functions:

- *Basic image parameters* record information crucial to displaying a viewable image. A total number of 30 metadata elements are defined, providing information on the file format including the applied compression method, colour space and the logical structure of the image file segments (either strips or tiles). Also, metadata elements on the file are stated, for instance image file identifier, image file location, file size and applied checksum method. Five metadata elements are mandatory and eight are mandatory if applicable. Figure 2.7 contains the metadata elements in the XML data format that are part of the ‘Format’ section.
- *Image creation* metadata elements record information important for understanding the technical environment in which a digital image file was captured. This category of the [NISOZ39.87:2002] data dictionary consists of 38 metadata elements. A number of metadata elements are intended to document features of the analogue source material scanned to create a digital still image. Also, data on the image production (for instance, host computer, operating system, model and settings of the digital capture device) is part of this category. According to [NISOZ39.87:2002] none of the metadata elements is mandatory, but for 15 of the metadata elements their usage is recommended.
- *Imaging performance assessment* metadata elements record information that allows evaluation of the quality of the digital image, or output accuracy. The operative principle in this section is to maintain the attributes of the image inherent to its quality. The 36 metadata elements in this category serve as metrics to assess the accuracy of output (today’s use) and of preservation techniques, particularly migration (future use). Four metadata elements are mandatory and 10 are mandatory if applicable. This category of the [NISOZ39.87:2002] data dictionary consists of three sections.

  The section ‘Spatial metrics’ deals with the grid of pixels of the digital image. The width and length of the digital image, and pixel resolution are part of this section. Also the width and height of the scanned object is covered in this section.

The section 'Energetics' provides data on the global energetic response and archiving space of the imaging device and subsequent digital file. The number of bits for each component for each pixel, and data on the coding of colour and grey levels are part of this section.

The third section, 'Targetdata', consists of metadata elements that can be used to benchmark the quality of the digital capture process with the help of either internal or external targets. The name and manufacturer of the target and data on the performance are also part of this section. The section also contains the most important metadata elements relevant for quality assessment of the digital image and a benchmarked conversion process.

– *Change history* metadata elements record information about the process applied to an image over its life-cycle. The seven data elements (of which three are mandatory if applicable) of this category are used to document the source, systems and settings used in all digital-to-digital operations that have occurred after the creation of the digital image.

The metadata elements of the [NISOZ39.87:2002] data dictionary build and expand on technical metadata available in other standards, such as the TIFF image file format version 6.0 [TIF92], TIFF/EP [ISO12234-2:2001] and DIG35 [DIG00].

Preservation metadata is documentation that plays a role in the long-term access of digital objects. The OAIS reference model (see section 2.2.2) can be used for the formulation of the metadata elements that users in the future, the Designated Community, need in order to understand and process the digital object. In general, metadata elements that are part of a metadata schema can be stored in three ways. In the first place metadata elements can be stored in the header of the image file, such as the information fields in a TIFF image file. Metadata elements can also be stored in the file system through the names of the directories and image files. In the third place metadata elements can be stored in a separate database.

The project 'Automatic exposure: Capturing metadata for digital still images' (see: [WAI04]) is working towards the automatic capture of technical metadata by digital imaging devices. The goal of the project is to lower the barrier for memory institutes to capture technical metadata elements for digital images. The project is based on the [NISOZ39.87:2002] set of metadata elements and was initiated by the Research Libraries Group (RLG).

The [NISOZ39.87:2002] data dictionary presents a comprehensive list of technical metadata elements relevant to the management of digital still images. In this context, management refers to the tasks and operations needed to support image quality assessment and image data processing throughout the image life-cycle. Quality assessment is defined broadly, as it refers to both machine operations and curatorial evaluations.

Technical metadata have been identified to 'anchor' meaningful attributes of image quality that can be measured objectively, such as detail, tone, colour and size. This standard frequently refers to images maintained in the TIFF digital image data format. The TIFF standard is a highly flexible and platform-independent format that is supported by numerous image-processing applications. The TIFF specification is publicly available to all users. The structure of the header includes a rich set of technical information important for long-term preservation such as for colorimetry, calibration, gamut tables, etc. The information is also very useful for remote sensing and multispectral applications. The repeated references to and examples citing the TIFF digital image format standard within the [NISOZ39.87:2002] data dictionary can be extended to other file formats. The [NISOZ39.87:2002] data dictionary indicates the information and metadata that all image files should contain as well as additional information related to digital image production.

EXIF and DCF

The EXIF standard [EXI02] and related DCF specification [DCF98] originate in the digital camera manufacturers community.[41] These standards are relevant for born-digital images. 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 at the beginning of the files und uses the standard colour space sRGB.[42]

Most digital cameras support the EXIF standard and DCF specification, for instance digital cameras manufactured by Canon, Kodak, Sony and Olympus. The EXIF format was 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 interchangeability, which will allow images captured on one camera to be viewed on another, or output directly to a printer.

The EXIF standard 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 information fields of the TIFF image file format [TIF92] with additional EXIF tags. Figure 2.4 contains the technical metadata according to the EXIF standard as created by a common consumer digital camera. The metadata is shown as an XML data file. The EXIF standard contains many more tags, but the camera does not use all of them. According to the EXIF

[41] Details on EXIF and DCF can be found at: <http://www.exif.org> [cited 24 August 2004].

[42] More information on the sRGB colour space can be found at: <http://www.w3.org/Graphics/Color/sRGB.html> [cited 15 August 2004].

```
<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 2.4 EXIF Metadata in XML format

standard, the code ‘1’ for the tag <ColorSpace> refers to the sRGB colour space.

As is illustrated by Figure 2.4, 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 of the three date tags are identical, as can be seen in Figure 2.4. Compressed image files are recorded as JPEG[43] image files with application marker segments inserted. Uncompressed files are recorded in TIFF version 6.0 format [TIF92]. Related attribute information for both compressed and uncompressed files is stored in the tag information format defined according to the TIFF image data standard. Information specific to the camera system and not defined in the TIFF image data format is stored in private tags registered for the EXIF format specification.

The fact that the EXIF standard 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.

[43] The JPEG compression method is defined in standard: [ISO10918-1:1994]. See: <http://www.jpeg.org/jpeg/index.html> [cited 14 August 2004].

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, such as printers. DCF is also applicable to products for writing image files on an interchangeable storage medium.

DCF consists of three specifications:

- *Media specification*. (Specifies state of data on a storage medium)
- *Writer specification*. (Specifies the recording function, for instance by a digital camera)
- *Reader specification*. (Specifies the playback function, for instance by a printer)

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. The importance of metadata for long-term access to digital objects is undisputed. A wide range of potentially relevant metadata schemas and related specifications do exist and the selection and application of the most relevant set of metadata elements depends on a number of factors, such as the tradition of the community of which the memory institute is part, the value and importance of the analogue original and the required quality of the digital surrogate.

Digital 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, even in the long term.

The Global Digital Format Registry (GDFR), as described in [ABR03] and [ABR04], is an example of an initiative that investigates the possibilities to establish a sustainable format registry.[44] The provisional data model for the GDFR includes properties of the registry itself and properties of the format. The format properties are subdivided into 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 is still far from being an operational production registry.

[44] More information on the GDFR and links to references can be found at: <http://hul.harvard.edu/gdfr/> [cited 12 May 2004].

The National Archives in the UK started a file format registry under the name PRONOM.[45] 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.

Besides Format Registries, tools have been developed to perform format-related 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 or metadata from an object. Whenever external metadata is submitted to a repository in connection with digital objects it should be checked for consistency.

JHOVE (JSTOR/Harvard Object Validation Environment) is an extensible framework for this format-related identification, validation and characterisation of digital objects.[46] The JHOVE programme 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 fit very well into the OAIS reference model, mainly related to the Producer and Archive entities.

The means of structuring the process of analysing the risks related to the probability that a digital file format can no longer be rendered in the future is covered in [LAW00]. The first steps towards the establishment of a methodology for investigating and measuring factors of digital formats and providing guidelines for preservation action plans can be found in [STA04].

2.2.4 Application profiles

Metadata supports the durability of objects, such as digital images. Digital images are alive as long as people and processes can use and interpret both the digital codes that represent the pixels of the digital image and the metadata that contains information about the digital image. The bitstream of a well-preserved digital image can probably be accessed even without metadata, but the chances are that it will be difficult to understand the meaning and purpose of the digital image. Several guidelines and standards exist that are relevant for the creation of metadata for digital images based on analogue historical sources (see Table 2.2). The fact that there are so many metadata schemas is an indication that several communities have different viewpoints and approaches to the creation and use of relevant metadata schemas. The diversity of metadata schemas is really a healthy sign be-

[45] The PRONOM System is accessible via <http://www.nationalarchives.gov.uk/pronom/> [cited 12 May 2004].

[46] The JHOVE software is made available publicly under the GNU General Public License (GPL) from the project website: <http://hul.harvard.edu/jhove> [cited 12 May 2004].

cause metadata schemas need to be fit for purpose and images tend to be highly heterogeneous.

As illustrated in section 2.2.3, a wide range of metadata schemas does exist and a number of Designated Communities are developing, using and maintaining these schemas. In practice, however, metadata schemas are often designed and applied in a pragmatic ad-hoc way. The metadata element 'title', for instance, can be interpreted in a number of ways. For a discussion of the semantic aspects of information brokering, see [KAS00] pp. 10-13. The interpretation of metadata elements that are part of a metadata schema can vary within different groups of users. Metadata elements are taken from existing metadata element sets and adapted for local use. People tend to 'mix and match' terms from multiple standards in order to meet the descriptive needs of a particular project or service. This 'mix and match' process is described by Heery and Patel [HEE00]. The concept 'namespace' is used to give terms that are part of metadata schemas an unambiguous reference and a unique identity.[47]

The set of metadata elements that are drawn from a number of metadata schemas combined and optimised for a particular local application is called an 'application profile'. By definition, an application profile cannot introduce new metadata elements. Each metadata element has to come from an existing metadata schema. Thus, application profiles reuse existing metadata elements. The difference between a metadata schema (or 'namespace schema', as it is called by Heery) and an application profile is that a metadata schema only declares metadata elements whereas an application profile reuses existing metadata elements. The Dublin Core Metadata Element Set [ISO15836:2003] is an example of a metadata schema. In the hypothetical situation in which all relevant schemas for all metadata requirements of all Designated Communities can be brought together, we still have to 'mix and match' the metadata elements in order to compile a set of metadata elements that cover all relevant issues and topics. This set of metadata elements is called an application profile.

An application profile has seven requirements (ref. [HEE00]). It declares which metadata elements from which namespace are used. It is possible to override default definitions. The source to which the application profile refers must be identified. Dependencies among vocabulary terms must be specified. It must be possible to define multiple entity classes. This means that, for instance, both people and resources can be described. Guidelines about the local use of particular terms must be provided (in free text) and the application profile should be able to express controlled vocabularies. The concept of application profiles is a good way to construct the most appropriate metadata schema for a specific purpose and audience.

[47] The 'namespace' concept is associated with the W3C Recommendation 'Namespaces in XML'. An XML namespace is defined as 'a collection of names, identified by a URI reference, which are used in XML documents as element types and attribute names'. See: <http://www.w3.org/TR/REC-xml-names/> [cited 22 April 2004].

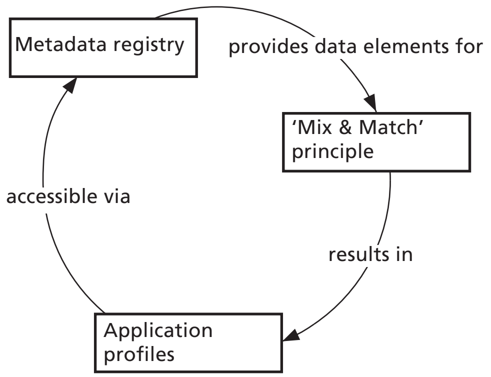


Figure 2.5 Relation between 'Mix & match' principle, Application profiles and Metadata registry

Registries of metadata elements

The length of time and the organisational effort it took for the library community to adhere to standardised principles for bibliographic metadata suggests that it will be virtually impossible to achieve in the short term a generally accepted and correctly applied set of metadata elements that enables the durability of digital objects, such as digital surrogates of historical photographs.[48] An alternative approach to formulating, maintaining and applying metadata elements relevant for the preservation of digital objects is the 'mixing and matching' of existing metadata elements and the registration of the metadata elements in metadata registries. This principle is illustrated in Figure 2.5.

Metadata elements have a name and associated semantics. This is implemented in a number of ways, resulting in a situation where metadata elements cannot be addressed unambiguously and this hampers interoperability. This is illustrated by an analysis of the attributes and properties of metadata elements, as they are part of a number of registries on the Web. The metadata elements of three registries are compared: the Dublin Core Metadata Registry,[49] the registry of the Renardus project[50] and the registry of the CORES project.[51]

[48] For a number of limited functions, generally accepted sets of metadata elements will be realised, such as the Dublin Core Metadata Element Set [ISO15836:2003], which is applied on a large scale for the realisation of 'resource discovery'. Also, for specific digital objects used by a well-administered user community the adherence to a standardised set is possible. The handling of electronic records by the archival community is an example of this.

[49] The Dublin Core Metadata Registry is accessible at: <http://www.dublincore.org/dcregistry/> [cited 14 August 2004].

[50] The Renardus application profile can be found at: <http://renardus.sub.uni-goettingen.de/renap/renap.html> [cited 14 August 2004].

[51] The web address of the Cores Registry is: <http://cores.dsd.sztaki.hu/> [cited 14 August 2004].

Table 2.3 Names of metadata elements that are part of three different metadata registries

Dublin Core registry	*Renardus registry*	*CORES registry*
Label	Name	ID
Definition	Label	Name
Description	Namespace	Definition
Is Defined By	DC Refinement(s)	Comment
RDF Type	R Refinement(s)	Data Type
Type	DC Encoding Scheme(s)	Obligation
Has Version	R Encoding Scheme(s)	Maximum Occurrence
Issued	Obligation	Refines
Modified	Repeatable	Element Set
Is Refined By	LQ 'LANG'	Annotations
Usage Example	DC Definition	Administrative metadata
	DC Comment	Element Usages
	R Definition	Refined By
	R Comment	
	Best Practice	
	Open Questions	

As stated in the text on the website of the Dublin Core Registry, 'The Dublin Core Registry is designed to promote the discovery and reuse of existing metadata definitions. It provides users and applications with an authoritative source of information about the Dublin Core element set and related vocabularies'. The Renardus project[52] created a portal to a number of European Subject gateways, mainly for academic resources. An application profile has been created to harmonise the metadata schemas that are used by the participating gateways. The central objective of the CORES project[53] is formulated as 'to encourage the sharing of metadata semantics. CORES will address the need to reach consensus on a data model for the ground-rules for declaring standard definitions of terms, as well as local usage and adaptations, will enable the diversity of existing standards to 'play together' in an integrated, machine-understandable Semantic Web environment'. More information on CORES can be found in [HEE03].

Table 2.3 contains an overview of the names of metadata elements of three metadata registries. The table demonstrates that metadata registries do not have a common way of expressing the features of metadata elements. The only attribute of a metadata element that is part of each of the three metadata registries is the attribute 'Definition'. The Dublin Core registry as well as the Renardus and CORES registries interpret the attribute 'Definition' in the same way. The definition of the

[52] The Renardus gateway system is available via: <http://www.renardus.org> [cited 14 August 2004].

[53] The CORES project website: <http://www.cores-eu.net> [cited 14 August 2004].

metadata element ‘Title’, for instance, is stated as ‘a name given to the resource’. The differences in the way the metadata registries formalise the attributes of metadata elements are illustrated by the fact that the Dublin Core registry uses the attribute ‘Label’ whereas the Renardus registry uses both ‘Name’ and ‘Label’ and the CORES registry uses the attribute ‘Label’ to designate the name of a metadata element.

A standardised way to formulate and express metadata elements is relevant for the creation of appropriate preservation metadata. This issue is covered in the second section of the next chapter by selecting and evaluating a standard that makes it possible to formulate metadata elements in an unambiguous way.

The realisation of a metadata registry is hampered by two factors. Metadata registries as structural, permanent services do not exist at the moment. Also, the procedure for compiling and fixing the properties of a metadata element according to a standard is not carried out on a large scale. Some small-scale prototypes do exist but have a rather experimental character. The first steps are for standards makers to agree on a common approach to formulating the metadata elements in their standards in a uniform way. Heery proposes defining application profiles in RDF [HEE03].

2.2.5 XML: Durable storage format for digital data?

This section addresses the relevance of the XML data format as a future-proof, non-proprietary data storage format. The main purpose of the Extensible Markup Language (XML) standard is to describe data by using special instructions called ‘tags’. XML version 1.0 was formally released in February 1998 ([BRA98] p. 8). It turned out to be a rapidly maturing technology with a wide range of powerful applications, mainly for the management, display and organisation of data. The XML text format standard was developed by the World Wide Web Consortium (W3C)[54] shaped by experience of previous markup languages, such as the ISO standard SGML.[55] The XML 1.0 recommendation[56] describes the characteristics of an XML document. A document that meets the requirements of the XML standards is called a ‘well-formed’ document.

Based on the XML data format, a number of related specifications have been developed that increase its applicability. The Document Type Definition (DTD) construct and the XML Schema recommendation can be used to define constraints on the logical structure of XML-formatted documents and to support the use of predefined storage units. For document formatting, several components of the Extensible Stylesheet Language (XSL) are available. Xpath, Xlink and Xpointer are advanced hyperlinking and addressing functions. Other members of the XML family

[54] More information on the XML-related activities of the W3C can be found at: <http://www.w3c.org/xml> [cited 27 September 2004].

[55] ISO 8879:1991 *Information processing – Text and office systems – Standard Generalised Markup Language* First edition. International Organisation for Standardisation.

[56] W3C Recommendations are similar to standards published by other organisations.

of standards are XML Query and Xforms. The Resource Description Framework (RDF) is a language for representing information about resources. RDF has an XML syntax.

There are two alternatives for expressing application profiles in XML format: RDF Schema and XML Schema. RDF Schema provides support for rich semantic description, while XML schema is best suited to supporting explicit, structural, cardinality and data typing constraints. As both structures are relatively new, there is no extended experience or knowledge available as yet.

A vehicle for expressing structure is not enough. The semantics of the elements and the way the elements are selected, used and adjusted are even more important. Of course, the Internet is the medium to reach Designated Communities and to communicate about schemas. But it should be noted that it is possible that some relevant key knowledge may not accessible via the Web.

Having on the one hand a means to compile a framework for gathering and disseminating specific metadata schemas – application profiles – and on the other hand ways to express the metadata elements – RDF Schema / XML Schema, there is nevertheless a gap between the two. This gap is caused by the problems of formulating the consensual, shared and formal description of the metadata elements that are important in a given domain for a given Designated Community. The recognition and identification of a Designated Community is not sufficient. In this respect the 'Semantic Web' initiative guided by the W3C consortium can be of importance. The goal of the Semantic Web is the creation of a semantic data structure on the Internet that can be understood by both humans and machines [FEN03].[57]

The exact requirements of application profiles as well as ways to express them are under discussion and under construction. The implementation of the principles has just started, so knowledge on its management, costs and usability by a Designated Community is not available. In the end, of course, users and stakeholders now and in the future will determine the relevance and success of the application profile of durable preservation metadata both about the analogue original and about the digital surrogate.

Durability issues of the XML data format

The durability issues of the XML data format are based on the following observations:

- *XML is self-describing*. An XML-formatted document consists of character data and mark-up. The mark-up encodes the description of the document's storage layout and logical structure and is human readable.
- *XML is a standard*. The World Wide Web Consortium (W3C) is developing interoperable technologies (specifications, guidelines, software and tools) to

[57] The main website of the Semantic Web initiative can be found at: <http://www.w3c.org/2001/sw/> [cited 18 August 2004].

lead the Web to its full potential.[58] XML is one of these technologies.

- *XML is media and application independent*. XML-formatted documents can be stored on any digital storage medium and are based on standard character set encoding sets, such as Unicode.
- *XML is license free*. No license is required to use the XML data format.
- *XML is both machine and human readable*. An XML document is a text document containing mark-up tags and characters that can be read by humans. Computer programs are available that process XML-formatted documents in a sensible way based on the mark-up tags in the documents.
- *No obligatory mark-up tags are required*. XML-related specifications, such as XML Schema and XSL, are used to fix the features of mark-up tags that are used in XML documents.

Resource Description Framework (RDF)

A data model is required for the identification and definition of data elements and application profiles. RDF and RDF Schema, developed by the W3C, are XML-based languages for describing resources, such as data elements and application profiles. In [HJE01] an overview of RDF is given. RDF provides a format for describing objects, but it does not say anything about what terms should be used to make the statements that describe the object and what those terms mean. In RDF the definition of the terms used in the assertions is given in a separate document called RDF Schema and in a specific language, the RDF Schema language. RDF data can be viewed in three representations: as a graph, as triples and as an-XML formatted text file. The representations contain data on the resource, the property and the value.

The relevance of RDF for the creation of metadata for digital preservation is described in more detail. With RDF it is possible to express the relations between the parts in the statement in the form of the triple: 'resource', 'property', 'value'. The property of an object can be another object. This makes it possible to formalise the statement. RDF is a framework for formalising assertions. It does not contain any vocabularies for authoring the metadata itself. Anyone can design new vocabularies, as long as they conform to the XML and RDF syntaxes. RDF defines the rules for defining and expressing semantics. It does not define the semantics itself for a topic in question. The elements that are part of an RDF statement are stored as namespaces. Hjelm defines namespaces as a mechanism for identifying names and making them unique ([HJE01] p. 42). The identification of a namespace is a URI (Uniform Resource Identifier) and is defined using a family of reserved attributes. As there is no registry, you have no way of knowing which schemas already exist.

[58] See <http://www.w3c.org> [cited 27 September 2004].

```
<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 2.6 Small part (of the 'File' section) of a METS document, stating that four digital images (in TIFF format) belong to the same group

Preservation metadata in XML format

The XML data format can be used to express preservation metadata in a number of ways. The Encoded Archival Description (EAD),[59] for instance, is a document type definition (DTD) expressed in the XML data format that declares elements that are based on the ISAD standard (see Table 2.2). The EAD is a standard for encoding archival finding aids, such as inventories, and is compatible with the data elements of the ISAD standard.

Another XML-based construct relevant for digital preservation is the Metadata Encoding and Transmission Standard (METS). The METS schema is a standard for encoding descriptive, administrative and structural metadata relating to objects within a digital library, expressed using the XML Schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress and was developed as an initiative of the Digital Library Federation.[60]

METS is intended as a method for storing and expressing metadata, especially for audiovisual sources and their digital representations. METS addresses the multiplicity of data element sets in recent years. In that sense it is a kind of meta-metadata or 'wrapper'. METS can be considered as an integral part of the OAIS reference model and serves to facilitate the exchange of digital objects from one repository to another. A METS-formatted document consists of seven major sections:

1. *METS header*: the header describes the METS document itself, for instance information on the creator.
2. *Descriptive metadata*: may point to descriptive metadata external to the METS document, for instance 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

[59] The DTD of the Encoded Archival Description can be found at: <http://www.loc.gov/ead/> [cited 19 July 2004].

[60] The main website of the METS initiative is: <http://www.loc.gov/standards/mets/> [cited 19 July 2004].

```
<?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 2.7 Data elements of the [NISOZ39.87:2002] standard expressed as the XML Schema 'MIX' and included in a METS wrapper

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 (contains a small part of the file section of a METS document.)
5. *Structural map:* the 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.[61] An XML Schema defines the allowable contents of an XML document. With METS, a collection of related digital objects, for instance the digitised pages of a book or digitised photographs from an album, can be joined together. METS has a liberal approach to the format of

[61] See: <http://www.w3c.org/XML/Schema> [cited 27 September 2004].

the metadata elements that describe the digital objects, as any format can be used. Systems supporting the METS standard are still at the prototype phase. Figure 2.6 contains an example of a small part of a METS document, namely of the File section. The example makes it clear that four digital images belong to each other. The four images could be digital raster images of a series. 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.

An XML Schema that contains the metadata elements of the data dictionary [NISOZ39.87:2002] is available. This XML Schema, 'NISO Metadata for Images in XML Schema' (MIX), provides a format for interchange and/or storage of technical metadata.[62] [DAL04] states that MIX is an extension schema of METS. Figure 2.7 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 elements. According to the [NISOZ39.87:2002] 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 ([ISO14721: 2003]), 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. 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 [NISOZ39.87:2002] data dictionary can be considered as a part of the administrative metadata of the METS specification. As an XML Schema implementation of the [NISOZ39.87:2002] data dictionary exists under the name 'MIX', the three standards (OAIS, METS and [NISOZ39.87:2002]), each relevant on a specific level for digital preservation of digital raster images, can be combined. This combination is illustrated in Figure 2.8. An XML Schema of the EXIF specification would also fit within the Administrative Metadata section of a METS standard.

2.2.6 Business models for digital preservation

This section discusses some issues related to costs that are intrinsic to the preservation of digital objects. Experience with digital preservation is recent and still evolving. A study by Wall [WAL03] concludes that very few institutes have a medium- to long-term plan relating to business model issues in the development of digital cultural content. Cultural issues influence the way digital preservation policy is implemented and how business models are developed. In a study by Weiss [WEI02], a European and a US policy regarding public sector information are distinguished. According to Weiss, the European approach promotes a cost recovery

[62] This XML Schema is referred to as 'NISO Metadata for Images in XML', and abbreviated as MIX. See: <http://www.loc.gov/standards/mix> [cited 27 September 2004].

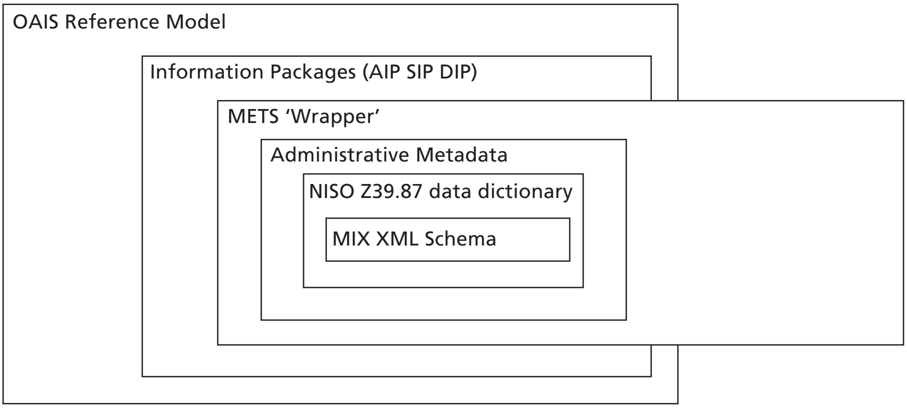


Figure 2.8 Relation between OAIS Reference Model, METS Schema and the [NISOZ39.87:2002] Data Dictionary

policy whereas the US stands for the open and free access policy ([WEI02] p. 2).

One of the first cost calculations for a digital archive was carried out in 2001 by Dürr and Van der Meer [DUR01]. The study makes clear that digital archiving is a costly activity. Metadata assignment, administration and quality control and other human activities turned out to be the major cost factors in the operation of a digital archive. Appraisal and selection of what should be kept for the long term is an important activity that can control the costs of digital archiving.

More recent studies concerning the costs of digital preservation focus on the storage costs for digital objects. A combination of factors influence storage cost in traditional and digital repositories. These include: the repository's unit rate for billing, the type and number of media being deposited, the number of versions being deposited, and the relationship between information content and media format. There is a significant gap among various content and media types. Research and development efforts to close this gap are essential to model, scale and sustain digital preservation services.

In [CHA03] Chapman compares costs in 2003 for the preservation of analogue formats on the one hand and the preservation of digital formats on the other hand. The study is based on practices at the depository service for analogue materials of Harvard University Library and the digital archiving service of OCLC (Online Computer Library Center). Both organisations manage centralised repositories optimised for long-term storage of library collections and fully recover their operational expenses by charging owners annual rates for managed storage services, regardless of materials use. A comparison is made between the costs to store comparable collections in analogue formats in the Harvard Depository and in digital format in the OCLC storage.

OCLC storage rates are $60 per gigabyte per year for 1 to 100 gigabytes of data, $32 per gigabyte per year for 101 to 1,000 gigabytes of data and $15 per gigabyte per year for more than 1,000 gigabytes of data. This price is for 'bit preservation' services only. These include data management and back-up, ongoing virus and fixity checks, periodic media refreshment, disaster recovery, and support of administrative tools for owners to update metadata and generate reports. Prices have not been set for 'full preservation' where OCLC would be obliged to provide standard bit preservation services, plus the capability to render intellectual content accurately, regardless of technology changes over time. OCLC accepts the following digital image formats: BMP, GIF, JPEG and TIFF. The cost for storage of analogue material at Harvard University is assessed per billable square foot (BSF). BSF represents a cubic dimension of 12' x 12' x 9'. Current rates are $3.91 per BSF per year for standard climate-controlled storage and $9.85 per BSF for film vault climate-controlled storage.

According to [CHA03], only for ASCII-formatted digital material is digital storage more cost effective than traditional storage of equivalent analogue material. The gap between costs for digital storage and analogue storage is most apparent for non-textual formats such as photographs. The storage costs of digital images are between 50 (for 35 mm film) and 200 times (for 4 cm. x 5 cm negative) more expensive than for analogue film storage. The storage costs for an uncompressed 230-megabyte TIFF image are $3.35 per year, whereas the managed storage of a high-resolution 4 x 5 cm analogue transparency costs only $0.016. Chapman mentions three solutions for closing this large digital-to-analogue gap: limit the size of the files by scanning at lower resolution, use compression, or keep the digital masters outside the repository but deposit a reduced-resolution uncompressed version. All solutions are advised against because they would lower the image quality too much and result in too much risk of losing the images. Chapman argues that providers should be persuaded to price repository storage in units other than object size, such as numbers of items or rates of access.

In 2004 Harvard University Library reported on a Digital Repository Service (DRS). The funding of the DRS of Harvard University is divided into three components with different sources of support ([CHA04] p. 99):

- The DRS *infrastructure* is financed through an annual billing algorithm based on the usage of the library's union catalogue. The DRS infrastructure costs include: development staffing, maintenance, operations and various costs for university information systems costs.
- DRS *storage costs* are recovered by charging the object owner fees, at the current rate of $5 per GByte per year. Compared with the 2003 OCLC rates stated above, the rates of Harvard University in 2004 are much lower.
- DRS *object transformation costs* will also be recovered by charging the object owner fees. These costs will vary depending on the type of transformation and the number of objects involved.

The deposit of objects is free just like the performance of maintenance on deposited objects and associated metadata by an administration tool. The metadata of the objects must be formatted in the XML data format according to a batch

DTD. Cornell University Library's Digital Consulting and Production Services (DCAP) [RIE04], established in 2003, is another example of an organisational infrastructure for ensuring the sustainability of digital collections. DCAP was established to 'provide a framework for the image creation and management process, including requirements analysis, implementation, assessment and archiving' ([RIE04] p. 191).

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 ([LEF02] p. 5). Another way to categorise inherent persistence is whether the materials are homogeneous. This means closely tied to known and consistent rules regarding structure, technical parameters and metadata.

It can be concluded that even at the creation stage of a digital object its relevance for long-term access, and thus its longevity, is an important factor. This observation is confirmed by several other studies and projects resulting in organisations that have an integrated view on the creation, use and archiving of digital objects. An example of this is the Australian VERS (Victorian Electronic Records Strategy) initiative (ref. [QUE04]) based on the concept of encapsulated digital objects.

Increasingly, memory institutes are developing specific procedures and services concerning the long-term archiving of digital objects. The National Library in the Netherlands, for instance, is carrying out a feasibility project aimed at the establishment of a national service dedicated to the long-term archiving of digital images in the TIFF image file format (ref. [VER04]). Another example is the classification of digital objects and their related digital preservation issues by the National Library of New Zealand (ref. [ROS03]).

This chapter contains a review and discussion of the state of the art of digitisation of analogue source material by memory institutes in general and of historical photographs in particular. Also, the current state of the art concerning the preservation of digital objects is described. The next chapter focuses on building blocks for the durability of digital surrogates of historical photographs by elaborating on a number of premises.

3 Durable digital surrogates of historical photographs

In contrast with chapter 2 where existing approaches concerning the digitisation of photographic items by memory institutes and the state of the art regarding digital preservation in memory institutes are described, this chapter brings together existing knowledge originating from a number of sources and not necessarily related to digital preservation. Based on the existing approaches as provided by literature and projects in this chapter, three premises are introduced and discussed. These premises, or statements that are assumed to be true, are formulated inductively and are considered as important building blocks for the longevity of digital surrogates of historical photographs. The clarification of the premises presented in this chapter contributes to the current state of the art concerning aspects relevant for digital durability.

Premises are arguments that lead to a conclusion. In this study building blocks that contribute to the durability of digital surrogates of historical photographs are presented and analysed. The intention of this chapter is to make the obvious link between premises and conclusions clear for the main stakeholders for the research presented in this study, namely memory institutes confronted with issues related to the long-term preservation of digital objects.

The first premise concerns the conversion of historical photographs into digital master images with the help of objective conversion quality metrics. The second premise addresses the unambiguous formulation of preservation metadata and the third premise deals with the persistent access to digital surrogates of historical photographs in institutional repositories.

The premises are now described in greater detail:

– *Premise one: Tools and standards are available for the creation of high-quality digital surrogates of historical photographs*. It is no longer necessary to call upon image scientists to realise an objective, standardised conversion process. his benchmarked conversion process is very important for the durability of the digital surrogate, because it creates a strong, controllable link between the analogue original and its digital surrogate. The creation of high-quality digital master images, enabled by this benchmarked conversion, is an important factor of digital longevity. Long-term access requires active attention and a justification for the allocation of resources for this lies in the assumption that high-quality digital surrogates of historical photographs deserve long-term attention.

- *Premise two: Preservation metadata can be formulated in an unambiguous way.* Preservation metadata are important for digital longevity. The application profile construct is the most appropriate structure to formulate preservation metadata. Preservation metadata consists of metadata elements that contain the information needed in order to access and understand the object that is described with the metadata elements. For the formulation and unambiguous definition of metadata elements, standards are available. These standards are important to strengthen the quality of application profiles that cater for the 'mix and match' principle. A number of metadata schemas are potentially relevant for the documentation of digital objects.
- *Premise three: Permanent access to digital objects can be realised by persistent identifiers, preservation repositories and open access.* The longevity of digital objects is strengthened if long-term access to these objects is guaranteed. A distributed networked environment based on open standards is an important factor in the realisation of a situation in which digital objects are stored in a durable way. It is assumed that digital longevity is not obstructed by the current state of the art in terms of storage media.

The next three sections of this chapter each discuss one of the premises in detail.

3.1 Benchmarked digital capture process

In this section the first premise 'Tools and standards are available for the creation of high quality digital surrogates of historical photographs' is examined in detail. The main purpose of a benchmarked digital capture process is to base the conversion of an analogue original into a digital surrogate on objective measurable image quality settings. In the context of this study, benchmarking relates the quality of a digital surrogate to the physical characteristics of the original historical photograph. The scan-once constraint requires that the quality of the digital surrogate be as high as possible. In this section the concept of image quality is more closely examined, followed by a discussion of tools and standards that are available to create faithful digital surrogates.

3.1.1 Image quality

Shaw makes clear that image quality issues have been studied for over a century by a number of scientific disciplines and that contemporary problems include the translation of techniques developed for the evaluation of analogue imaging processes into the digital domain [SHA99]. Often, astronomy has provided substantial contributions to the fundamental understanding of image quality issues.

According to Engeldrum, no widely accepted formal or de-facto definition of image quality exists. He proposes using the following definition: *Image quality is* *the integrated perception of the overall degree of excellence of an image* ([ENG04A] p. 160).[63] Image quality is an outcome of many complex processes that may involve, among other things, software algorithms, chemistry, physics and the psychology of human judgement ([ENG04A] p. 161). In [ENG99] and [ENG04B] Engeldrum introduces the 'Image Quality Circle' (IQC) to address the lack of a structure or framework for the complete understanding of image quality. The IQC facilitates the establishment of image quality goals and specifications and a structure for putting image quality into products. The primary focus of the IQC is commercial imaging where image quality is cast as a 'beauty contest' selection from images produced by competing products. The IQC consists of four interrelated elements, with three links that connect two of the elements:

- *Customer Image Quality Rating*. This is the overall image quality rating as judged by customers, using appropriate psychometric methods. This quality element of the IQC is not relevant for digital surrogates of historical photographs, because the image quality of a digital surrogate depends on the extent to which features of the analogue original are present in its digital surrogate, acting as a digital master. For the quality judgement of the derivatives that are based on the digital master, the Customer Image Quality Rating is relevant. This element of the IQC can also play a role in the quality assessment of the analogue originals and can be included in a selection procedure for analogue sources that are qualified for a benchmarked digital conversion.
- *Technology Variables*. These are the items that can be manipulated to change the image quality, such as dots per inch, paper thickness of a printed picture and number of megapixels in the image sensor. The list of Technology Variables is almost endless and is of limited longevity. New technology will almost certainly raise questions regarding the usefulness and validity of today's Technology Variables. The other two elements of the IQC, Physical Image Parameters and Customer Perceptions (see below), address the longer term and provide a comprehensive framework for image quality.
  - *System / Image models* are formulas, physical models, algorithms, or computer code that connect the Physical Image Parameters and the Technology Variables. They predict the Physical Image Parameters from the Technology Variables.
- *Physical Image Parameters* are the quantitative functions and parameters, such as modulation transfer, density and colour. They are obtained by physically measuring the image with instruments or computations on an image file and have historically been called objective measures of image quality. Physical Im-

[63] In this study an 'image' refers to historical photography and its digital surrogates. Engeldrum uses the term image in a much broader way and defines it as: '... a colorant arranged in a manner to convey 'information' to a human observer. Colorant is used in its most generic sense. It can be ink, plastic (toner), wax, dye, silver, phosphors, light emitters, and so on. The function of the image is to visually communicate information.' ([ENG04B] p. 448)

age Parameters can be any measurable aspect of an image. A large number of Physical Image Parameters have been developed by the imaging industry and formalised in international standards. In the next section, standards and tools acting as Physical Image Parameters are examined.

  - Visual Algorithms connect Physical Image Parameters to Customer Perceptions. These algorithms compute the value of a 'ness', for instance sharpness (see below). Visual Algorithms have an extensive history in photographic image quality and their use has been extended to digital imaging.

- *Customer Perceptions (The 'Nesses').* The major perceptual attributes of image quality are dimensions like darkness, sharpness and graininess. They form the basis of the quality preference or judgement of an image. Most perceptual attributes are visual. An example of a non-visual perceptual attribute of quality is the quality of the substrate upon which the images are printed. For the quality of digital images, non-visual perceptual attributes are very relevant, such as the quality of the storage medium and the quality of the image file format. Understanding the Customer Perceptions is considered as the key to understanding the IQC.
  - *Image Quality Models* (IQM) connect Customer Perceptions with Customer Image Quality Rating. The purpose and function of an IQM is to predict the image quality judgement from the value of the 'nesses' in the image. Image quality judgements can be made with respect to a reference, such as a reference image or a standardised image system.

The Image Quality Circle is a complete and comprehensible model for the assessment of image quality. Almost half of the elements of the IQC, from Visual Algorithms to the Customer Image Quality Rating, require human judgements, making psychometric scaling an important aspect for image quality.

The research in this study is aimed at the quality of the digital image, acting as a surrogate of an analogue original. Thus, image quality is interpreted in line with the Physical Image Parameter element of the IQC. Therefore, the emphasis will be on objective measurements based on standardised procedures. The human judgement factor related to quality is only in an indirect sense relevant for the benchmarked conversion of historical photographs; that is for the quality assessment of the analogue original image prior to digitisation and the features of the digital surrogates that are based on the digital master images. The remainder of this section will examine the creation of faithful digital surrogates by controlling the Physical Image Parameters.

3.1.2 International standards on digital imaging performance metrics

The two most evident settings of digital capture devices are related to the tonal depth and the spatial resolution of the output image. The tonal depth is expressed as the number of bits per pixel. A minimum of 8 bits per pixel is the norm for the digital capture of greyscale material and a minimum of 8 bits per colour channel is the norm for the digital capture of colours. Within the preservation community, for a couple of years already the consensus has been that digital master images should contain more than 8 bits per channel ([FRE97C] p. 114). The spatial resolution determines the number of pixels in the digital image and is specified by the scanning resolution for flatbed scanners expressed in 'dots per inch' (dpi) and for digital cameras expressed as pixel raster size. In principle, all interfaces of digital capture devices are able to cope with these two settings. A problem, however, is that the ways the digital capture device compiles the pixels of a digital image based on the two settings very much depend on the specific features of the device. The quality of the optics and the scanning mechanism of the digital capture device affect the characteristics of the resulting digital image. The precision of the photosensitive cells determines the accuracy of the encoding of the intensity and wavelength for each point on the image ([TAN98] p. 139).

Several digital capture devices apply interpolation techniques, resulting in 'artificial' pixels in the output image. To a certain extent each digital capture device is a 'black box'. This is a problem because some aspects of the digital capture workflow are hidden and are not under control by the operator of the conversion process. In *Measuring quality of digital masters* [FRE00A], Frey describes the issues that influence the quality of digital master images. Frey states that some of the solutions for the creation of digital images based on objective quality parameters are still under development or not yet commercially available. Since the publication of Frey in 2000 the situation has improved in the sense that both standards and methods have become available that can be used by memory institutes to create digital masters with an objective predictable quality. Also, metadata schemas have been developed that can be used to store this type of preservation metadata. It is the purpose of this chapter section to discuss these standards and methods.

Specific calibrated test charts, or targets, are required for the benchmarking of digital capture devices. The standardised targets have specific normalised features, such as grey and colour patches and fine-grained patterns. The quality of digital capture devices can be measured by relating the features of the targets in digital form to the physical characteristics of the original targets. This characterisation by means of quality metrics is an important instrument for creating digital surrogates with a predictive quality. The image of any source can be linked back to that source with appropriate capture documentation and benchmarking targets. While the original source characteristics are not unequivocally recoverable, suitably accurate reconstructions of the source can, in principle, occur ([NISOZ39.87:2002] p. 28). Targets are used as concise physical benchmarks for absolute energetic and spatial information about the item of interest at the time of capture. They are, in essence, Rosetta stones for the source. As such, their utility is undisputed whenever corrections or faithful reconstructions of the source document are required. Depending on workflows and philosophy, targets can be considered as either external or internal to a digital image. Internal targets are part of a digital image by being within the field of view at the time of capture. External targets are typically captured session-to-session and usually give temporally sparse information between image captures. For stable capture environments their utility can be equivalent to that of internal targets. Since they are not part of the digital image itself, their location must be managed in order to maintain a link to the source ([NISOZ39.87:2002] p. 38).

For both reflective and transparent photographic material, flatbed scanners and digital cameras can be used. Transparent photographic material can be digitised by capturing the light that is sent through the photograph, and the digitisation of reflective photographs requires a device that captures light projected on to the source. In this section, both types of digital capture device and both types of photographic material are used in their most general way. Scanners, film, displays and image processing can all be characterised through performance metrics that quantify the way in which their functions modify light intensities and their associated spatial structure.

One of the first digital capture projects that came up with a rather simple method for the benchmarked capture of sources for an archival project is the digital capture project of the UC Berkeley Library [JON02]. Based on published sources, a protocol for capture resolution, master file format and other settings was developed. For the capture of tonal information, recommendations were difficult to find, so the Library project developed its own procedures. With the help of greyscale patches the tonal response of the digital capture device is measured, then ideal aimpoints are calculated from the digitised patches. The settings of the digital capture device are adjusted until the measured values match the calculated aimpoints. The images are coded according to the RGB colour space. Now the digital capture device is calibrated and the digital master images have a standardised tonal range. This method is described in [JON02] and has resulted in standardised digital master images. Other descriptions of practical tests concerning the quality performance of digital capture devices can be found in [LOE01] and [WUE01].

Frey mentions four digital image quality parameters: tone reproduction, detail and edge reproduction, level of image noise, and colour reproduction ([FRE97A] p. 598). Recently, standards, targets, tools and methods have become available for the assessment of these parameters. They are described in the next sections of this chapter.

Assessment of tone reproduction

'Tone reproduction is the most important aspect of image quality' ([MRP04] p. 4). Objective assessment of optical density is an important Physical Image Parameter. The tone reproduction of a digital capture device can be measured by the opto-electronic conversion function (OECF). The OECF defines the relationship between the exposure, or reflectance, and digital count value of a digital capture device. The OECF is the single most important link between the physical photometric legacy of the analogue source object and its digital counterpart as a digital image file. The OECF allows one to evaluate the effective gamma[64] applied to an image, any unusual tonal manipulations and device non-linearity ([WIL03] p. 78). The OECF represents the relationship between the optical input values of a digital capture device and its digital output image values. ISO standards are available that allow for performance evaluation in a common and unified metric. These standards and tools will help to characterise the equipment. They will not fix problems.

Test methods for measuring the OECF of a digital capture device are described in [ISO14524:2000]. This ISO standard describes in detail how the OECF of a digital capture device can be determined. In the standard, the required standardised test target is described as well as the number of times the chart shall be imaged and how the OECF can be presented in tabular and/or graphical form. The OECF of a digital capture device expressed as a graph reveals to what extent the device can distinguish density values.

The hardware of a capture device behaves linearly with the amount of detected light. The data from the capture device is processed by the software driver and ultimately controls the behaviour of the OECF through gamma, shadow, highlight and contrast functions. The driver of the capture device contains a number of non-standardised functions in terms of controlling the OECF of a scanner, such as gamma, highlights and contrast.[65] The OECF can be created in three steps. First, a greyscale step target with known optical reflectance is scanned.[66] The documentation that comes with the target contains information on the densities[67] and reflectance of each patch. Also, software tools are available that contain information on the densities of the patches, such as specific Photoshop plug-ins. In the second

[64] Gamma here means the degree to which a digital capture device is non-linear in tonal behaviour, represented as the exponent of a power function. Most imaging devices that measure light have linear responses: doubling the light intensity doubles the brightness. The human visual system is non-linear: when the light intensity is doubled, you don't see the light as twice as bright. To relate the linear responses of digital capture devices to human perception, a logarithmic scale has to be applied.

[65] A higher gamma setting at capture will increase the brightness of a displayed image. For maximum and consistent image content capture, these functions should be disabled (i.e. shadow=0, highlight=255, auto-contrast=off).

[66] Targets can be obtained from companies like Applied Image Group, see: <http://www.appliedimage.com> [cited 12 September 2004]. Also the Kodak calibration targets can be used, such as Q-13 and Q-60. Analysis software tools are available from the website of the International Imaging Industry Association: <http://www.i3a.org/downloads.html> [cited 12 September 2004]. The tools 'sfrmat' and 'sfrwin' analyse both the tone and detail reproduction (see next section) based on a slanted edge target.

[67] Density is the degree to which a surface absorbs light. Density is a logarithmic value. Based on the known density, the reflectance of each patch can be calculated. The much-used Kodak Q-13 reflective Greyscale target contains 20 patches from white to black. Three patches have a code: A, M and B, with the corresponding densities 0.0, 0.70 and 1.60, which represent average highlight, middle tone and shadow values in colour and black-and-white reflection copy. With the help of a densitometer, the density values of a specific target can be verified and adjusted.

step the average count values (for instance, the RGB values of the pixels of a patch) are computed. In a third step the density and reflectance values are paired with the average count values, for instance in a table or graph. This table or graph provides a reversible connection between the photometric features of the original analogue source (represented by a standardised step target) and the way the digital capture device is able to discriminate between grey levels and the RGB or other values these grey levels have in the digital image.

It is important to document the OECF for ease of future image producing and rendering (see: [JON02]). The [NISOZ39.87:2002] data dictionary contains a number of data elements that can be used to store the OECF of a digital capture device. These data elements are: 8.2.5 ('GrayResponseCurve', the optical density of each possible pixel value), 8.2.6 ('GrayResponseUnit', enumerated list determining the precision of the information contained in the GrayResponseCurve (for instance, 1 = tenths of a unit)) and data element 8.3.4 ('PerformanceData', identifies the path of the file that contains the image performance data relative to the target identified in 8.3.2). Metadata on the targets used to calculate the OECF can be created with the data elements with the code 8.3.2. Here, the manufacturer, the name, type, number and other features of the target can be expressed.

To a large extent good colour management relies on maintaining grey levels at reproduction. The individual RGB colour channels of a grey-scale target will not have identical OECF values. By documenting the OECF, any mismatch can be corrected. In the OECF a sufficient count value buffer should be taken into consideration for future tone and spatial image manipulation. An alternative to documenting the OECF is to include an image of a standardised greyscale target in each individual image file and to maintain it as a part of the image file. The OECF is a prerequisite for facilitating image reproduction and for image quality and colour management tasks.

Assessment of detail and edge reproduction

The ability of a digital capture device to distinguish the relative contrast of finely spaced detail on the original is an important image quality issue. Related quantitative metrics, or Physical Image Parameters, are the number of pixels, the sampling frequency (dpi, ppi), the resolving power of the digital capture device, the Spatial Frequency Response (SFR) and the Modulation Transfer Function (MTF). The resolution specifications given by the manufacturers of digital capture devices are ambiguous, mainly because different sampling and interpolation techniques are applied. Thus, the output image of a digital capture device has to be analysed in a standardised way in order to determine to what extent the device is able to reproduce the details of an original.

Several optical, mechanical and digital filtering imaging mechanisms influence the actual resolving power of a digital capture device. The resolution is determined by the highest frequency at which light and dark parts of an image are distinguishable. The apparent sharpness of a displayed digital image can be improved by using a digital sharpening filter but at the expense of increased image noise. To capture all information in an analogue continuous signal, it should be sampled at a rate that is equal to or higher than twice its highest frequency. This highest frequency is sometimes referred to as the Nyquist Frequency ([JAC00] p. 409). The Nyquist Frequency represents the highest frequency that can be faithfully reconstructed.

The limiting resolution is the spatial resolution at which finely spaced features are no longer visually detectable. Traditionally, the visual inspection of a digitised target is used to measure the limiting resolution. This method has a number of drawbacks. It relies on human readable subjective evaluation, the results are target contrast dependent, and it provides little image quality insight information.

For a number of types of digital capture devices, standards are available to determine to what extent details and edges can be reproduced. For the assessment of still picture cameras, [ISO12233:2000] is available. For print scanners, [ISO16067-1:2003] can be used and [ISO16067-2:2004] is dedicated to film scanners. ISO resolution test charts are defined in these standards. The test charts include patterns with fine details, such as edges, lines, square waves and sine wave patterns. The standard contains a description of the purpose of each test pattern on the test chart as well as the conditions under which the test chart should be digitised. Also, the reporting of the outcomes of the image analysis is specified in the standard.

With the help of a computer algorithm the digital image of the test chart can be evaluated.[68] The software calculates the Spatial Frequency Response (SFR) of the imaging system. SFR is the measured amplitude response of an imaging system as a function of relative spatial frequency ([MRP04] p. 20). The SFR is an important Physical Image Parameter indicating the image quality. It is an indicator of contrast loss as a function of spatial frequency. The SFR is also known as the Modulation Transfer Function (MTF). The SFR is measured by capturing an image of a slanted edge target at scanner settings of interest. A digital image of a good slanted edge target is required. The SFR or MTF is a descriptive plot that measures the extent to which image detail contrast (modulation) is maintained by an imaging component or system. The SFR provides a way to analyse the influence of imaging components on the retention and reproduction of image detail and sharpness. The principles of the measurement of the Modulation Transfer Function are described in ISO standard [ISO15529:1999].

The 'TargetData' section[69] of the [NISOZ39.87:2002] data dictionary contains data elements that can be used to document the results of the detail and edge reproduction assessment. Whether the target is internal or external is documented in data element 8.3.1 ('TargetType'). The location of the external targets is man-

[68] The software 'sfrmat' and 'sfrwin' can be used to evaluate the quality of the captured standardised target and can be found in the ISO section at: <http://www.i3a.org/downloads.html> [cited 12 October 2004].

[69] The 'TargetData' section is part of the 'Imaging Performance Assessment' category of functions, as described on pages 63 of this study.

aged by data element 8.3.3 ('ImageData'). Detailed information on the characteristics of the target can be stored in the data elements of section 8.3.2 ('TargetID'). A reference to the file that contains the image performance data relevant for the target can be stored in data element 8.3.4 ('PerformanceData'). [NISOZ39.87:2002] does not prescribe which image quality metrics should be calculated. For this, the ISO standards [ISO12233:2000], [ISO16067-1:2003] and [ISO16067-2:2004] can be used.

Assessment of image noise

Image noise is a general term applied to error, or unwanted fluctuations, in images. It can have many sources and several forms, such as random noise due to film grain or low exposure to a detector or compression artefacts. Image noise is related to the reliability of the detected signal and is influenced by the gamma settings. Correlated noise is noise that bears some regularity, such as streaks and patterns. Non-correlated noise is a random distortion and much more difficult to detect with the human eye ([GAN99] pp. 229-234).

The occurrence of image noise is related to the dynamic range setting of the digital capture device. Increasing the dynamic range, for instance, from 8 to 12 bits in principle indicates that the number of tone levels increases from 256 to 4096 levels, also leading to a much higher maximum optical density. But together with the higher bit level more image noise is introduced and this influences the quality of the digital image. Thus, image noise limits the effective bit depth of a digital capture device. Williams suggests an alternative definition of dynamic range to cater for this phenomenon: 'The dynamic range is the extent of energy over which a digital capture device can *reliably detect* signals, reported as either a normalised ratio or in equivalent optical density units' (ref. [WIL03] p. 79). The stronger the signal the better, but the level of noise in the signal should be as low as possible.

Two ISO standards are associated with the assessment of image noise: [ISO21550: 2003] for the calculation of the dynamic range for digital film scanners and [ISO15739: 2003] for the noise measurement for electronic still picture cameras. Both standards use identical techniques for characterising dynamic range and noise. This method is described in [BUR01]. With the image noise assessment methods described in the ISO standards, the signal-to-noise (S/N) ratio can be determined. For this a target has to be digitised under controlled conditions.[70]

The calculation of the incremental S/N ratio consists of three parts and requires sophisticated statistical operations. First, the calibrated patches on the target make it possible to determine the relation between densities and the associated digital count value. An incremental signal is calculated, which means that from a range of optical densities all average related digital count values are calculated. An

[70] The ISO 15739 Noise Test Chart Utility can be downloaded from: <http://www.i3a.org/downloads.html> [cited 18 September 2004]. This utility contains a spreadsheet that quantifies the noise created by a digital capture device at different dynamic range values.

incremental signal indicates how well a device can detect small density differences, on average. Secondly, for the calculation of the noise, the standard deviation of the count level associated with a particular density level is calculated. In the third step the outcomes from the first two steps yield the incremental S/N ratio.

Quantifying the influence of image noise on the real dynamic range of a digital capture device is not an easy procedure, mainly because of the application of advanced statistical calculations. But the S/N ratio is a very relevant Physical Image Parameter. An 8 bit per pixel scanner in theory has an optical density of 2.4 (and this is what manufacturers claim in product literature), but an actual density of about 1.5 is probably much more realistic.[71]

Metadata concerning the assessment of the image noise of a digital capture device can be stored in the [NISOZ39.87:2002] data dictionary. Section 8.3 ('TargetData') consists of a number of data elements that can be used for this purpose. A path to the file that contains information on the image noise performance can be stored in data element 8.3.4 ('PerformanceData').

Assessment of colour reproduction

The fourth Physical Image Parameter that can be used to measure the image quality concerns the faithful reproduction of the colour information of the analogue original in its digital surrogate. As a lot of historical photographic material is monochromatic, the capture of colour information can be irrelevant. This depends on the rendering intent as covered in section 2.1.4 of this study. For the digital capture of monochromatic photographic material, the accurate conversion of the grey tones is sufficient.[72] Comprehensive coverage of colour reproduction issues can be found in [BER00].

The assessment of accurate colour registration by digital capture devices very much resembles the way in which the assessment of tone reproduction is achieved. Calibrated targets containing colour patches are digitised, and tools and methods are used to analyse the digitised targets in order to assess the quality of the digital capture device. Before the standards and tools to assess colour are discussed, two issues are clarified. First, the rationale for a device-independent colour coding method is covered, followed by a discussion of the physical colour registration limitations that digital input and output devices have. Both aspects are important for accurate colour management.

[71] By using digital capture devices with 10 bits per pixel or higher the density limitations are largely abolished. The density of the analogue original is a factor to be taken into consideration, because the tone reproduction in the digital surrogate must contain all tone levels of the original. As reflective photographs do not have a density higher than 2.0 (ref. [GAN99] p. 217), a 10 bits per pixel scanner will be able to capture the dynamic range of the original. Of course, the scanner has to be benchmarked and the results of the benchmark process have to be documented in order to create durable digital surrogates.

[72] Grey tones are the result of uniform RGB colour channels, so even in a 'black-and-white environment' colour channels are adjusted. See also 'Assessment of the tone reproduction' in this section.

The fundamental basis for all colour registration and reproduction is the three-channel design of the human retina – the nerve cells of the eye receptive to light. Three types of colour sensors in the retina are each responsive to different regions of the visible spectrum that correspond roughly to reds, greens and blues. Phosphors on digital displays and pigments on paper use the three colour channels to reproduce other colours. Additive colour mixing uses the colours red, green and blue and is also known as the RGB colour space. Monitors and scanners are RGB devices. Subtractive colour mixing uses the three colours cyan, magenta and yellow in the CMY colour space. Printers use this method for the creation of colour.[73] RGB and CMYK are device-related colour models, because the values are related to the features of the device that produces the colour.[74] This means that the same colour values will produce different colours on different devices and that, in order to produce the same colour on different devices, the colour values have to be changed.

The solution to this problem is to use a device-independent colour model. The *Commission Internationale de l'Eclairage* (CIE) has developed a number of device-independent colour models that attempt to model the way humans perceive colour.[75] The device-independent CIELAB colour model is often used in situations where the representation of colour must meet high standards. CIELAB values have to be correlated with device-dependent colour models like RGB and CMYK in order to make colours. CIELAB codes give device-dependent colour codes an unambiguous meaning.[76]

All output devices such as printers and monitors have a fixed range of colours that they can reproduce. This range of realisable colours is called the colour gamut. Digital input devices have a maximum dynamic range, as discussed earlier in this chapter. The differences between device gamut and dynamic range complicate accurate colour representation. In the 1980s and early 1990s many companies developed colour management systems (CMS) that enabled colour consistency, but these systems were not compatible with the equipment of other companies. One of the goals of the International Colour Consortium (ICC), an industry consortium founded in 1993, is to define an open format for the specification of colour that all vendors can use and that enables consistent colour data exchange between

[73] Besides Cyan, Magenta and Yellow, printers also use Black ink in order to get better results. This is the reason why printers use the CMYK colour space, where K represents the Black colour component.

[74] The colour Green, for instance, can be defined in a number of RGB values; for example, the values [R:57 G:68 B:21], [R:56 G:68 B:23] and [R:56 G:67: B:20] can all be classified as 'Green'.

[75] The CIELAB colour space uses three primaries, called L*, A* and B*. L* represents the lightness, A* represents how red or green a colour is, and B* represents how blue or yellow the colour is. CIELAB represents all colours a human can see.

[76] A few years ago a large number of publications suggested using the simple and robust sRGB colour space for optimal rendering of images across platforms and devices. See, for instance, [FRE97C]. The 'extensible' RGB standard ([IEC61966-2-2:2003]) contains guidelines for transformations of sRGB to richer colour spaces such as CIELAB.

devices from different vendors. A CMS based on ICC standards consists of four basic components:

- *Profiles*. A file that contains enough information to let a CMS convert colours into or out of a specific colour space, such as the RGB or CMYK colour spaces. The ICC profile is specified in [ICC04].
- *Profile Connection Space (PCS).* The colour space used as the intermediate form for conversions from the source profile to the destination profile.
- *The Colour Management Module (CMM).* This is the software, or 'engine', that performs calculations needed to convert the RGB or CMYK values that are contained in the profiles.
- *Rendering intents.*[77] There are four methods that deal with the out-of-gamut problems.[78] These methods are described in ([ICC04] pp. 8-13). For digital images that have to be colorimetrically true, two rendering intents are relevant: the *relative colorimetric intent* (the colours of the surrogate match the colours of the original as closely as possible) and the *absolute colorimetric intent* (the colours are reproduced as exactly as possible, independent of output media and viewing conditions).

The work done by the ICC facilitates the standardised, consistent description of the behaviour of digital image devices and this is important for the quality of the digital surrogate created with the equipment. A wide range of devices and software are 'ICC compatible' and often have good user interfaces. Profiling, or description of behaviour, is not enough for achieving high-quality durable digital images. The behaviour of the digital image devices has to be optimised for the benchmarked conversion of a specific analogue original. This means that the device has to be calibrated.

For the creation of durable digital master images, a controlled digital capture or input process is important. A good input profile will inform the CMS what colours the capture device can distinguish. The creation of an input profile for devices, such as scanners and digital cameras, means that RGB values registered by the device are compared with measurements in a device-independent colour space, such as CIELAB. Physical targets are required, for which measurements of the colour patches on the target are available in a description file. The most common targets used are the IT8.7/1 (transparent) and IT8.7/2 (reflective) targets.[79] Software tools are available to calculate the profile based on the image of the target.

The most appropriate ISO standard regarding the encoding of colour information is [ISO22028-1:2004]. The standard specifies the colour space encoding, the viewing conditions, the image state and reference medium. The standard also

[77] In section 2.1.4 of this study 'rendering intent' has another meaning and is defined as the basic principle of converting a photographic original into a digital surrogate.

[78] The four rendering intents distinguished by the ICC are: saturation, perceptual, absolute and relative.

[79] The Kodak Q-60 target is based on the IT8 standard.

specifies a set of requirements to be met by colour encoding. The colour management architecture currently in place allows the communication of consistent colour across all applications, devices and operating systems.

The [NISOZ39.87:2002] data dictionary contains a number of data elements that facilitate the creation of metadata relevant for the assessment of the quality of the colour representation of a digital surrogate. The colour space of the image data is stored in data element 6.1.4.1 ('ColorSpace'). The name and location of the profile is stored in data element 6.1.4.2.1 ('ProfileName') and data element 6.1.4.2.2 ('ProfileURL').

3.1.3 Faithful digital surrogates

The relevance of image quality parameters has been recognised for a considerable time, but only recently have standards and tools become available that can be used to benchmark the digital capture process of photographic material in an archival environment. The sensible application of the tools and standards requires some dedication, but it is no longer necessary to call upon image scientists to benchmark digital capture devices. On the basis of the outcomes of the benchmarking process, informed decisions can be made, such as the adjustment of settings, the re-scanning of a batch, or the replacement of equipment and software.

Also, constructs are available for the storage of the Physical Image Parameters in the form of preservation metadata. Preservation metadata is important for the durability of digital objects. The Research Libraries Group (RLG) has begun the 'automatic exposure' initiative to encourage scanning system manufacturers and imaging software vendors to populate formatted raster, still image objects with relevant metadata to the fullest extent possible (see [WAI04]). This initiative was inspired by the [NISOZ39.87:2002] data dictionary.

Some issues hinder the smooth and easy application of the performance metrics. Most of the standards are not freely available and must be ordered from standardisation organisations. A number of standardisation organisations are involved in the creation and maintenance of digital imaging performance metrics, such as ISO, IEC, ANSI, NISO and CIE, and not all activities are coordinated with each other. Within the standard documents, terms are not used consistently. Print scanners and reflective scanners, for instance, are the same. Obtaining the appropriate targets also can be complicated, as a number of different organisations produce them. Just as temperature can be expressed in a number of ways, image quality metrics also have a number of relevant units. Some of these are related to inches while others are related to the metric system. For the expression of detail, for instance, both line pairs per millimetre and cycles per millimetre are used. This may lead to complicated conversion calculations and confusion.

Figure 2.2 illustrates the durable digitisation process of photographic material. Published sources of best practice recommendations are available to translate the assessed digitisation requirements into digital imaging settings, such as scanning resolution and tonal range, and benchmarking based on standards is an important issue with regard to the establishment of digital conversion settings (see, for instance, [BEN02]). The dynamic range of a digital capture device should meet or exceed the dynamic range of the analogue original. A dynamic range of 8 bits per pixel for greyscale material and 24 bits for colour material is sufficient for good-quality output, but on the input side a higher initial tonal resolution is required to capture the full density range of the original and to compensate for image noise. In particular, transparent materials such as photographic negatives require a high bit depth during capture.

The measurement of the performance of a digital capture device is standardised for a number of features. The quality of the large area image capture can be measured by calculating the OECF. Capture of image detail can be measured by calculating the SFR and MTF. Also, image noise and artefacts can be measured, for instance, by determining the dynamic range of the capture device. A CMS based on the standards of the ICC enable the device-independent capture of colour information. Understanding the technology used helps interpretation of the results.

Benchmarking is required to ensure the usability, persistence and interoperability of digital objects. An important objective of benchmarking is to define baseline levels of quality that minimise or eliminate the need to digitise a work more than once. Faithful digital surrogates meet these criteria and are intended to render the original source accurately, with respect to its completeness, and the appearance of the original, including tonality and colour. Faithful digital surrogates will support the production of legible printed facsimiles when produced in the same size as the original. Ultimately, legibility and fidelity are subjective factors and related to the Customer Image Quality Rating.

3.2 Unambiguous formulation of preservation metadata

In this section, the premise ‘Preservation metadata can be formulated in an unambiguous way’ is examined. Metadata is important for long-term access to digital objects. A number of metadata schemas exist that can be used to support the durability of digital objects, such as digital surrogates of historical photographs. The OAIS reference model ([ISO14721:2003], see section 2.2.2) can be used to assess the role and value of the metadata elements that are part of the metadata schemas.

Two observations can be made concerning the selection and application of preservation metadata. In the first place, people tend to ‘mix and match’ metadata elements from a number of metadata schemas. The application profile construct can be considered as a solution for the compilation of metadata schemas that are most suited to a specific situation (see section 2.2.4). In the second place, it can be observed that, for the formulation and specification of metadata elements, a number of methods and structures do exist. The fixing of the semantics of metadata schemas is carried out in several ways, and this makes it difficult to use in an unambiguous manner the metadata elements of an application profile that are derived from a number of metadata schemas. Common issues that play a role in the evaluation and application of metadata elements are that they can have the same name in different metadata schemas but have a different meaning, or that metadata elements of different metadata schemas have different names but share the same semantics. The goal of this section is to provide a means of expressing metadata elements in an unambiguous way.

Metadata and library science

Libraries have a long tradition regarding the creation of documentation or metadata of information entities such as printed books and journals. The rules used to describe information entities are relevant for the creation of preservation metadata. This section covers relevant issues regarding the theory of description. This theory is required to define entities and metadata elements that form unambiguous preservation metadata.

The book *The intellectual foundation of information organization* [SVE00] states that bibliographic languages are required to organise and manage information. An important distinction is made between information and its embodiments. The computer revolution changed the nature of entities to be organised and the means of their organisation. One of the new problems relates to the nature of digital documents. A traditional document, like a book, tends to be coincident with a discrete physical object. A digital document can be unstable, dynamic and without identifiable boundaries. 'What is difficult to identify is difficult to describe and therefore difficult to organise' ([SVE00] p. 13). This certainly holds true for the wide range of digital objects that came into existence in the computer age.

According to Svenonius, the oldest and most enduring source of problems that frustrate the work of bibliographic control is the language used in attempting to access information. In a perfectly orderly language, each thing has only one name, and one name is used to refer to each individual thing. The construction of an unambiguous language of description, defined as 'a language that imposes system and method on natural language and at the same time allows users to find what they want by names they know', turned out to be very difficult to accomplish.

The language used to make descriptions is a bibliographic language, a special-purpose language that is designed and applied in accordance with a special set of rules. It is an artificial language. A description is a 'statement of the properties of a thing or its relations to other things serving to identify it' ([SVE00] p. 53). A 'work' is defined as particular disembodied information content. A 'document' is defined as material embodiment of information ([SVE00] p. 9). Since medieval times, in western culture authorship has been the primary identifying attribute of works, but perceptions of authorial functions have changed over time ([SVE00] p. 43). A work language describes information in terms of its attributes, such as author, title, edition and subject. A document language describes attributes that are specific to particular manifestations of works, such as publication attributes, physical attributes and location attributes. The rules for the design and application of a bibliographic language are contained in codes and standards. There are hundreds of these, for instance the AACR2R.[80] In the Anglo-American cataloguing tradition, it is a cardinal rule that the starting point for bibliographic description should be the physical form of the item in hand. Applying this rule results in a one-to-one relationship between documents and the descriptions that are their surrogates ([SVE00] p. 108). An ideal bibliographic record would be so full and accurate that a user would be able to determine the relevance of the document described without physically examining it ([SVE00] p. 117).

Two trends appear to be dominating current research and development. One is the increasing formalisation of information organisation as an object of study through modelling, linguistic conceptualisation, definitional analysis of theoretical constructs and empirical research. The second is the increasing reach of automation to develop new means to achieve the traditional bibliographic objectives, to design intelligent search engines, and to aid in the work of cataloguing and classification ([SVE00] p. 193). Concerning the creation and use of documentation, library science makes clear that the formulation and application of preservation metadata is complicated by the following factors:

- In the digital world the distinction between 'works' and 'documents' as distinctive items is not always clear. As photographs in the analogue world are often called 'non-book' items, this issue will be even more difficult for the creation of metadata for digital surrogates of historical photographs.
- Rich syntax rules (for arranging elements into statements) of a metadata language will enhance the quality of the metadata. But complicated syntax rules can lead to difficulties in implementing solutions.
- The semantics of bibliographic language elements can lead to problems. Persons and systems can misunderstand the meaning of language elements.

The complications mentioned above can result in ambiguously formulated preservation metadata.

Standardisation

The quest for a method to express metadata schemas in an unambiguous standardised way starts with an analysis of standardisation in general. This analysis will make clear whether a standard does exist that facilitates the creation of metadata elements. A standard is an agreed-upon way of doing something, a uniform set of measures, agreements, conditions or specifications between parties. Standardisation can be defined as the activity of establishing, with regard to actual or potential problems, provisions for common and repeated use, aimed at achieving the optimum degree of order in a given context ([SPI01] p. 1).

Standards developing organisations (SDOs) are formal and structured organisations that achieve consensus on a broad number of subjects.[81] An important

[80] AACR2R: 'Anglo American Cataloguing Rules, 2nd Edition'.

[81] The USA has about 50 private-sector standards developing organisations with active, ongoing standards-development programmes ([SPI01] p. 134).

international SDO is the International Organisation for Standardisation (ISO). Before some relevant ISO standards are discussed regarding the formalisation of metadata schemas, the standards development process in IT in general is covered.

Within IT, standards are a business issue. While other industries use standards as part of their product cycle, in the IT industry standards are used as a product differentiator ([SPI01] p. 258). The IT industry creates standards not for the end-users but for the intermediate layer of application providers and manufacturers. IT standards are part of the active consciousness of the end-users, who may not understand them, but who do, at times, demand them.

Several IT standards are developed by informal organisations such as consortia and alliances. De-facto marketplace-driven standardisation is important for IT standards. A de-facto standard is based upon market-driving skills such as product availability, product profile and timing. In the hardware sector the informal standardisation process has worked well. Devices from a multitude of different vendors can be used on a wide range of computer systems. Within the software sector, however, a lot of non-standardised products and services have been developed. There have been many attempts to create de-facto standards in order to achieve market dominance, but most of these have been unsuccessful. A few, however, have been very successful. If an organisation could not cause a de-facto standard to emerge, the second drive was to produce a standardised specification faster than the SDOs could. Often this is done as a consortium, a collection of like-minded companies. A consortium provides a formalised structure for accomplishing a set task within a defined time frame. Only competent marketing firms with technical capability succeed with de-facto standards over the long term, as the maintenance and growth of a de-facto standard are cumbersome ([SPI01] p. 259).

The informal manner of creating de-facto standards, the fact that maintenance of IT standards is troublesome and the tendency for IT standards to be replaced by new ones are indicators of the risk that standards in IT organisations do not last and that the standards are supported by only a certain number of the IT organisations. For this reason, it is relevant to look more closely to the output of a formal SDO such as ISO. The formal standards of ISO are probably more robust and durable, but it should be kept in mind that several organisations prefer the faster, informal way of working in consortia and alliances.

3.2.1 ISO Standards relevant for unambiguous formulation of data elements

An obvious place to look for prescriptive initiatives or meta languages to create standardised, high-quality metadata is ISO. This is a worldwide federation of national standards bodies from more than 140 countries, one from each country. ISO is a non-governmental organisation established in 1947. The mission of ISO is to promote the development of standardisation and related activities in the world with a view to facilitating the international exchange of goods and services, and to developing cooperation in the spheres of intellectual, scientific, technological and economic activity. ISO's work results in international agreements that are published as international standards. ISO defines standards as 'documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics, to ensure that materials, products, processes and services are fit for their purpose'.

The scope of ISO is not limited to any particular branch; it covers all technical fields except electrical and electronic engineering standards, which are the responsibility of the International Electrotechnical Commission (IEC). A joint ISO/IEC technical committee carries out work in the field of information technology. This joint technical committee, established in 1987, is known as 'JTC 1'. The need for a standard is usually expressed by an industry sector, which communicates this need to a national member body. The latter proposes the new work item to ISO as a whole. Once the need for an international standard has been recognised and formally agreed, the first phase involves definition of the technical scope of the future standard. This phase is usually carried out in working groups that comprise technical experts from countries interested in the subject matter. In the second phase, the consensus-building phase, countries negotiate on the detailed specifications within the standard. The final phase comprises the formal approval of the resulting draft international standard, following the publication of the agreed text as an ISO international standard.

By the end of 2004, ISO's work has resulted in some 12,000 international standards, representing more than 300,000 pages. The technical work of ISO is highly decentralised, carried out in a hierarchy of technical committees, subcommittees and working groups. In these committees, qualified representatives of industry, research institutes, government authorities, consumer bodies and international organisations from all over the world come together as equal partners in the resolution of global standardisation problems. Some 30,000 experts participate in meetings each year.

In order to find out whether ISO standards are available that enable the creation of robust, unambiguous and clearly defined metadata schemas, the work of relevant Technical Committees or subcommittees has been analysed. According to the ISO website, there are 224 Technical Committees.[82] Based on their title and scope of operation four of these committees are potentially relevant for the compilation of durable metadata schemas:

- The first obvious one – as this study uses ICT products – is the ISO/IEC joint committee (*ISO/IEC JTC1*) on 'standardisation in the field of information technology'.
- The second one is *TC 37* concerning 'standardisation of principles, methods and applications relating to terminology and other language resources'.

[82] The website of the International Organisation of Standardisation can be found at: <http://www.iso.ch> [cited 14 May 2004].

- The third Technical Committee of potential interest is *TC 46*, ‘standardisation of practices relating to libraries, documentation and information centres, indexing and abstracting services, archives, information science and publishing’.[83]
- The fourth Technical Committee is *TC 171*, ‘Document management applications’. The scope of TC 171 is ‘Standardisation of quality control and integrity maintenance in the field of document management’. Documents may be managed in micrographic or electronic form.

TC 37 and TC 46 are mentioned by Veltman as the way to ‘arrive at the ontological meaning of a term’ ([VEL01] p. 163). After analysing the work carried out by the Technical Committees as well as their subcommittees and working groups, the most relevant for the standardised creation of metadata schemas are the results of Workgroup 2, ‘Metadata’, of Subcommittee 32 ‘Data management and Interchange’, part of ISO/IEC JTC1. The most important output of ISO/IEC JTC 1/SC 32/WG 2 is the ISO/IEC 11179 family of standards. This group of standards addresses the specification and standardisation of metadata schemas and will be discussed in detail later in this dissertation.

For four reasons, the work of the committee working on the ISO/IEC 11179 family of standards has been selected as the most relevant result of a formal standardisation organisation for the unambiguous creation of preservation metadata schemas:

- The ISO/IEC 11179 family of standards is built upon normative documents created by other committees of the ISO/IEC community. Standard ISO 704, ‘Terminology work – Principles and methods’ [ISO704:2000], created by Technical Committee 37, for instance, is mentioned as a normative reference for the ISO/IEC 11179 family of standards.
- The committee that formulates the ISO/IEC 11179 family of standards (ISO/IEC JTC 1/SC 32 ‘Data management and interchange’) is very often mentioned as an important reference platform for other initiatives. In the draft business plan of ISO/TC 46 ‘Information and documentation’, for instance, a formal liaison with JTC 1/ SC 32 is presented ([ISO02] p. 4).
- The ISO/IEC 11179 family of standards is the most relevant initiative of the ‘metadata interchange standards list’ of the ‘Diffuse’ project. Diffuse is an EU-funded single, value-added, neutral entry point to up-to-date reference and guidance information on available and emerging standards and specifications that facilitate the electronic exchange of information.[84]

[83] Subcommittee 11, ‘Archives/records management’, of Technical Committee 46 has developed an ISO standard aimed at metadata for records and is potentially relevant for the creation of preservation metadata for digital objects. The first part of the standard, ‘Principles’, is available as [ISO23081-1:2004] whereas the more explanatory parts are not yet available.

[84] The website of the Diffuse project can be found at: <http://www.diffuse.org> [cited 13 January 2003]. Since 31 January 2003 the information on the Diffuse website has not been updated. No decision regarding maintenance of the service has yet been made. An archived

- The most important reason, however, is that ISO/IEC 11179 is mentioned as a reference for the standardised formulation of data elements that are part of specific metadata schemas. Each element of the Dublin Core Metadata element set [ISO15836:2003], for instance, is defined using a set of ten attributes from the ISO/IEC 11179 standard for the description of data elements.[85] Several initiatives to create registries of metadata schemas are based on the ISO/IEC 11179 standard.[86] Heery refers to ISO/IEC 11179 in relation to the development of metadata schema registries ([HEE02] p. 2).

3.2.2 ISO/IEC 11179 family of standards on specification of data elements and metadata registries

Among the numerous standards supported by ISO as a formal standardisation body, the ISO/IEC 11179 family of standards facilitates the robust, durable formulation of metadata. The quality of the metadata and the level of interoperability when using the metadata are important for the long-term access to digital objects that are described by the metadata. The ISO/IEC 11179 family of standards promotes the consistent representation of metadata, the consistent interpretation of metadata and the sharing of metadata. The ISO/IEC 11179 family of standards describes a model for classifying, naming, identifying and registering information in a way that promotes the understanding of information.

The purpose of ISO/IEC 11179 is 'to give concrete guidance on the formulation and maintenance of descriptions and semantic content that shall be used to formulate Data Elements or Value Domains in a consistent, understandable manner. It primarily does this by giving guidance for building, establishing and maintaining a Metadata Registry' ([ISO11179-1:1999] p. VI). A Data Element is defined as 'a unit of data for which the definition, identification, representation and permissible values are specified by means of a set of attributes'. ISO/IEC 11179 enables the correct, confident and unambiguous formulation and interpretation of Data Elements. The ISO/IEC 11179 family of standards consists of six parts:

- *Part 1: Framework*. The framework provides the context for associating the six individual parts of the ISO/IEC 11179 family of standards and is the foundation for conceptual understanding of Data Elements. The first edition of this

version of the website of the Diffuse project can be found at: <http://web.archive.org/web/20031229131742/http://www.diffuse.org/> [cited 15 June 2004].

[85] The reference description of the Dublin Core element set according to ISO/IEC 11179 can be found at: <http://www.dublincore.org/documents/1999/07/02/dces/> [cited 12 July 2003]. The revised reference description of the Dublin Core element set published in 2003 no longer referred the ISO/IEC 11179 family of standards, but does not exclude its use either.

[86] Metadata registries based on the ISO/IEC 11179 family of standards are: the ebXML repository ('electronic business using extensible markup language'), see: <http://www.ebxml.org> [cited 27 September 2004], the CORES registry system available at: < http://cores.dsd.sztaki.hu/> [cited 27 September 2004] and the METAPRO system from the US Environmental Protection Agency: <http://www.epa.gov/metapro> [cited 27 September 2004].

part of the standard was published in December 1999. It states that the basic components for open information processing are hardware, software, communications and data. Standards have been proposed for three (hardware, software and communications) of the four components and ISO/IEC 11179 will standardise the fourth component (data) ([ISO11179-1:1999] p. 10).

- *Part 2: Classification*. This part of the ISO/IEC 11179 family of standards provides a conceptual model for managing classification schemes. A classification scheme is defined as an: 'arrangement or division of objects into groups based on characteristics that the objects have in common, for instance, origin, structure, application and function ([ISO11179-2:2000] p. 5)'. Types of classification schemes are: keywords, thesauri, taxonomies and ontologies. There are many structures for organising classification schemes and many subject-matter areas have been described by classification schemes.
- *Part 3: Basic attributes of Data Elements (1994) / Registry metamodel and basic attributes (2003).* In 1994 the first version of this part of the ISO/IEC 11179 family of standards was published [ISO11179-1:1994] and in 2003 a new edition was published [ISO11179-1:2003]. The 1994 version, entitled 'Basic attributes of Data Elements', specifies a set of mandatory metadata items that a registration applicant shall provide for each Data Element. In addition, a list of potential additional items for use as needed is provided. Detailed characteristics of each basic attribute are given.

  The 2003 version, entitled 'Registry metamodel and basic attributes', specifies a conceptual model for a Metadata Registry. The 1994 version of the ISO/IEC 11179-3 supported only Data Elements. The 2003 edition of the standard also supports other metadata items associated with Data Elements, such as Data Element Concepts, Conceptual Domains, Classification Schemes, Value Domains and other related classes, called Administered Items. Table 3.3 contains an overview of the ten types of Administered Items as distinguished by the ISO/IEC 11179 family of standards.

  The increased use of data processing and electronic data interchange relies heavily on accurate, reliable, controllable and verifiable data recorded in databases. One of the prerequisites for correct and proper use and interpretation of data is that both users and owners of data have a common understanding of the meaning and representation of the Administered Items. To guarantee a common view of these items, a number of basic attributes are defined in part 3 of the ISO/IEC 11179 family of standards.
- *Part 4: Formulation of data definitions*. Part 4 of the ISO/IEC 11179 family of standards [ISO11179-5:1995] provides guidance on how to develop unambiguous data definitions. A number of specific rules and guidelines are presented that specify exactly how a data definition should be formulated. A precise and well-formed definition is one of the most critical requirements for a shared understanding of an Administered Item. Well-formed definitions are impera-

tive for the exchange of information. Only if every user has a common and exact understanding of the data can it be exchanged in a trouble-free way. According to ISO/IEC 11179-4, a data definition shall be unique, be stated in the singular, state what the concept is (not what it is not), be stated as a descriptive phrase or sentence, contain only commonly understood abbreviations, and be expressed without embedding definitions of other Data Elements or underlying concepts.

- *Part 5: Naming and identification*. Part 5 of the ISO/IEC 11179 family of standards [ISO11179-5:1995] provides guidance for the identification of Administered Items. Identification is a broad term for designating, or identifying, a particular data item. Identification can be accomplished in various ways, depending upon the use of the identifier. Identification includes the assignment of numerical identifiers that have no inherent meanings for humans. Names are semantic, natural language labels given to data items, and variations of these labels serve different functions. Some names are for human usage and comprehension; some names are for use in a particular physical system environment. Names are often user established and vary from one user to the other. The principles in [ISO11179-5:1995] describe the various functions of names and how names are used. One and only one identifier is required for each Administered Item within a registration authority. The identifier does not change as long as the Administered Item remains unchanged. Identifiers are unique only within a registration authority. Along with each name for a data item, the context must be provided.
- *Part 6: Registration*. Part 6 of the ISO/IEC 11179 family of standards [ISO11179-6:1997] provides instructions on how a registration applicant may register a data item with a central Registration Authority (RA) and the allocation of unique identifiers for each data item. Maintenance of Administered Items already registered is also specified in this part of the ISO/IEC 11179 family of standards.
  The uniqueness of a registered data item is determined by the combination of the RA Identifier, the unique identifier assigned to the item within an RA, and the version. These are also included in widely available metadata registries. An RA to which data items logically and functionally belong maintains each registry. The registries should be indexed and constructed so that those designing applications or messages can ascertain easily whether a suitable data

Table 3.1 Status of the ISO/IEC 11179 family of standards in August 2004

Parts of ISO/IEC 11179 family of standards	*ISO/IEC standard in:*	*Revision state August 2004*[87]
11179-1 Framework	1999	Final Committee Draft (May 2003)
11179-2 Classification	2000	Working Draft (March 2003)
11179-3 Basic attributes	1994	International Standard (February 2003)
11179-4 Formulation of Data definitions	1995	International Standard (July 2004)
11179-5 Naming & Identification principles	1995	International Standard (July 2004)
11179-6 Registration of Data elements	1997	Committee Draft (March 2003)

item already exists. Where it is established that a new data item is essential, the procedure should encourage its derivation from an existing entry with appropriate modifications, thus avoiding unnecessary variations in the way similar data items are constructed. This is in line with the 'application profile' concept (see section 2.2.4).

Registration will also allow two or more data items serving an identical function to be identified and, more importantly, it will identify situations where similar or identical names are in use for data items that are significantly different in one or more respects.

Registration is more complex than a binary status simply indicating whether a data item is either registered or not. Although it is tempting to insist that only 'good' data may be registered, this is not practical. Therefore, improvement in the quality of registered data is divided into levels (called registration status). In addition, there are status levels for administration between each of these quality levels. Collectively, these status levels are called administrative status. They indicate the point in the registration life-cycle currently attained by a registered data representation. This part of the ISO/IEC 11179 standard can be considered as a reference towards the several Data Element registry initiatives that currently exist.

[87] The revision process of a standard consists of several stages. The first stage is the 'working draft', followed by several successive 'committee drafts'. The last one is the 'final committee draft'. This work is carried out at 'subcommittee level'. The 'final draft international standard' (FDIS) is the last step before a document (under the normal development process) is approved as an International Standard. FDIS documents are balloted and approved at the 'technical committee' level

The ISO/IEC 11179 family of standards underwent three main editorial revisions.[88] The first edition, published in the period 1994–2000, focused mainly on the standardisation of Data Elements. The title of the first edition of the ISO/IEC 11179 family of standards was *Information technology – Specification and standardisation of data elements*. The creation of the second edition started in 2003 and is foreseen to end in 2005. The advance of the Internet as the dominant medium for electronic communication is the main force behind the revisions of the standard. The title of the ISO/IEC 11179 family of standards was changed to *Information technology – Metadata Registries*. This second edition extends the focus of the standard from Data Elements to Administered Items in general and defines a metamodel for a metadata registry. The development of the third edition of the standard has already started, and will result in published standards in the period 2006–2008. The third edition will extend the use of the XML data format further. illustrates the revision status of the ISO/IEC 11179 family of standards in August 2004.[89]

Recommendations and practices for registering Data Elements are described in technical report ISO/IEC 20943-1 *Information technology – Procedures for achieving metadata registry (MDR) content consistency. Part 1: Data elements* [ISO20943-1:2003]. Recommendations and practices for registering Value Domains are described in technical report ISO/IEC 20943-3 *Information technology – Procedures for achieving metadata registry content consistency. Part 3: Value domains* [ISO20943-3:2004].[90] Both technical reports are consulted later in this study in an experiment to examine whether preservation metadata can be created in compliance with the ISO/IEC 11179 family of standards.

ISO/IEC 11179-3 Registry metamodel and basic attributes

Whereas the focus of the first version of the ISO/IEC 11179 family of standards was on the management and formulation of basic attributes of Data Elements, in the second version the management of Administered Items and the registry metamodel is the central issue of the standard. A metadata registry (MDR) is an information system for registering metadata. A metadata register is the information store or database maintained by a metadata registry. Part 3 of the ISO/IEC 11179 family of standards dealing with the basic attributes is the most important part of the standard and was the first part that was revised.

According to the [ISO11179-3:2003] standard, a metamodel is a model that describes other models. A metamodel provides a mechanism for understanding

[88] Information on the most recent state of the ISO/IEC 11179 family of standards can be gathered by monitoring the document library of the ISO/IEC workgroup available from the World Wide Web: <http://metadata-stds.org/Document-library> [cited 15 September 2004].

[89] Source: The website of ISO/IEC JTC1 SC32 WG2: <http://metadata-stds.org> [cited 5 August 2004].

[90] The following parts of the ISO/IEC 20943 standard are under construction: Part 2: XML structured data and Part 4: Overview [cited 4 October 2004].

Table 3.2 Ten types of Administered Items

	Type of Administered Item	*Description of Administered Item*
1	Classification Scheme	The descriptive information for an arrangement or division of objects into groups based on characteristics, which have the objects in common.
2	Conceptual Domain	A set of value meanings, this is the semantic content of a value.
3	Context for the Administered Item	A universe of discourse for which a name or definition is used.
4	Data Element	A unit of data for which the definition, identification, representation and permissible values are specified by means of a set of attributes.
5	Data Element Concept	A concept that can be represented in the form of a data element, described independently of any particular representation.
6	Derivation Rule	The logical, mathematical and/or other operations specifying derivation.[91]
7	Object Class	A set of ideas, abstractions, or things in the real world that are identified with explicit boundaries and meaning and whose properties and behaviour follow the same rules.
8	Property	A characteristic common to all members of an object class.
9	Representation Class	The classification of types of representations, for instance 'amount' or 'number'.
10	Value Domain	A set of permissible values.[92]

the precise structure and components of the specified models, which are needed for the successful sharing of the models by users and/or software facilities. A metamodel is used to describe the structure of a metadata registry. The registry in turn will be used to describe and model other data. The registry metamodel is specified as a conceptual data model, for instance one that describes how relevant information is structured in the natural world; it is how the human mind is accustomed to thinking of the information ([ISO11179-3:2003] p. 27).

The metamodel need not be physically implemented as specified. An implementer can use the metamodel to develop a more specific logical data model that describes the same data, but as structured in an information system. A logical data model can be used directly for database design.

[91]The Derivation Rule as a type of Administered Item is not part of the first version of the [ISO11179-3:2003] standard. It is introduced in Technical Corrigendum 1 of ISO/IEC 11179-3:2003, published 1 April 2004.

[92] In 2000 ISO/IEC Technical Report 15452 on specification of data value domains was published. The report covers the identification, specification, development and reuse of enumerated, bounded data value domains for data elements (see: [ISO15452:2000]). The contents of this technical report are integrated with minor adjustments in the ISO/IEC 11179-3 standard published in 2003 ([ISO11179-3:2003]).

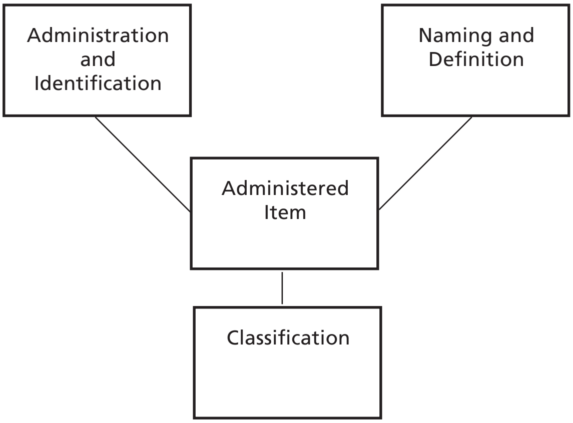


Figure 3.1 Common facilities for all Administered Items

In order to avoid confusion, the [ISO11179-3:2003] standard deliberately uses different terms for the specification of the metamodel construct and metadata objects prescribed by the metamodel itself. The registry metamodel is specified using a subset of the Unified Modelling Language (UML, ref. [ISO19501:2002]). It is important to distinguish between types of data and metadata on the one hand, and instances of these types and their associated values on the other hand. The metamodel specifies types of classes, attributes and relationships. Ten types of Administered Items are distinguished by the standard. Additional types may be defined as extensions to the standard. Table 3.2 contains an overview of the types of Administered Items.

Administered Items are identified once only and administered as single items within the registry. Administered Items are named and defined in at least one context. Context is defined by the [ISO11179-3:2003] standard as 'a universe of discourse in which a name of definition is used'. Administered Items may be classified in zero or more Classification Schemes. Figure 3.1 illustrates the relationship of the three common facilities to the Administered Items in the registry. The 'Administration and Identification' area specifies issues such as the organisation that registered the item, the organisation that submitted the item, the organisation that is responsible for the item, the dates the Administered Item was created, changed or became effective and the identifier of the Administered Item. The 'Naming and Definition' area is used to manage the names and definitions of Administered Items and the contexts for the names. It is recognised that an Administered Item may have many names that will vary depending on issues such as discipline, locality and technology.

The 'Classification' area provides a facility for registering and administering Classification Schemes and their constituent classification scheme items. A Classification Scheme is a terminological system intended to classify objects.

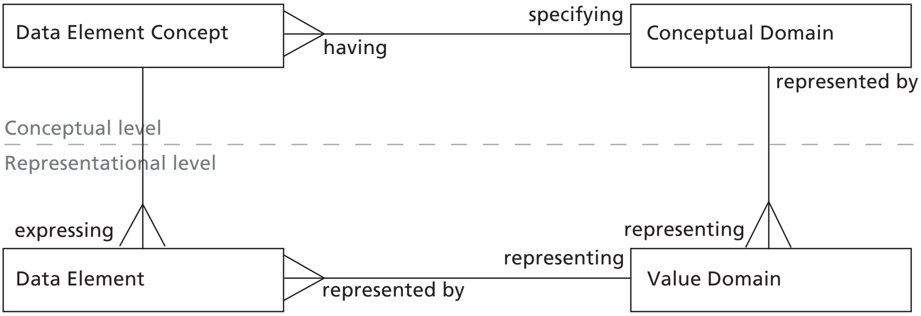


Figure 3.2 High-level metamodel of ISO/IEC 11179-3:2003

Four specific types of Administered Items are described in detail in the [ISO11179-3:2003] standard: Data Element Concept, Conceptual Domain, Value Domain and Data Element. These four types of Administered Items form the core of the ISO/IEC 11179 standard. The relationship between the four core Administered Items is illustrated in Figure 3.2. All Administered Items carry their own Administration Record information, allowing them to be identified, named, defined and optionally classified within a Classification Scheme.[93]

The purpose of the Data Element Concept is to maintain the information on the concepts upon which the Data Elements are developed. The concepts concentrate on the semantics and are independent of any internal or external physical representation. Conceptual Domains support Data Element Concepts and Value Domains support Data Elements. A Conceptual Domain is a set of value meanings, which may be enumerated or expressed via a description. The Value Domain is one of the key components of a representation. The Data Element provides the formal representations for some information (such as a fact, a proposition, an observation) about some concrete or abstract thing. Data Elements are reusable and shareable representations of Data Element Concepts. A Data Element is formed when a Data Element Concept is assigned a representation.

The Conceptual Domain and the Data Element Concept relate to the way the human mind perceives objects from the outside world. The Data Element and Value Domain relate to the way information on objects is represented. The formulation and expression of Data Elements and Value Domains are the most important parts of a metadata registry. For the adequate formulation of preservation metada-

[93] An example can illustrate the components of the metamodel. All the countries in the world can be considered as a Conceptual Domain. The country of birth of a European citizen is an example of a Data Element Concept. All European citizens have a country of birth. The name of a country is a Data Element. A Value Domain, such as the ISO 3166 standard list of country names and codes, can represent the name of a country. In this list the code 'NL' is used for the country with the English name 'Netherlands'. The code 'NLD' is also used in the ISO 3166 standard for this country.

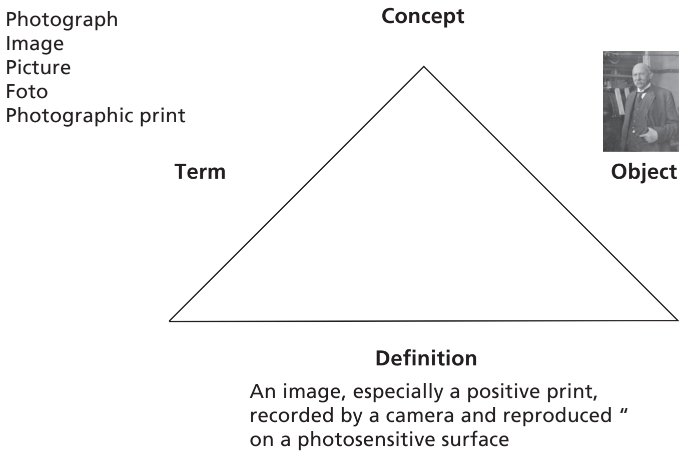

Photograph
Image
Picture
Foto
Photographic print
Concept
Term
Object
Definition
An image, especially a positive print, recorded by a camera and reproduced “ on a photosensitive surface

Figure 3.3 Model of a concept according to ISO/IEC 11179 family of standards

ta, the representational level of the ISO/IEC 11179 family of standards is the most relevant. The expression and representation of the conceptual level depends on the quality and relevance of the Data Elements and Value Domain.

A concept is defined as a ‘unit of knowledge created by a unique combination of characteristics’ ([ISO11179-3:2003] p. 10). A concept is captured with a definition and identifiers are assigned to uniquely identify concepts. This is illustrated in Figure 3.3. There are potentially many terms associated with any concept and many concepts associated with any one term. Each term has a context from which it is extracted. In Figure 3.3, for instance, the context of the term ‘Photograph’ is that it is the common name of the object. The context of the term ‘Foto’ is the Dutch translation and the context of the term ‘Picture’ is the broader term for the common name for the object.

The formalisation of the terminology used in the ISO/IEC 11179 standard is based on ISO 704 standard *Terminology work – Principles and methods* [ISO704:2000]. The main issues of this ISO standard are given. In the theory of terminology, concepts are mental constructs, units of thought, or units of knowledge created by a unique combination of characteristics. Concepts are organised or grouped by common elements, called characteristics. Essential characteristics are necessary and sufficient for identifying a concept. Other characteristics are inessential. The sum of characteristics that constitute a concept is called its intension. The set of objects a concept refers to is its extension. In natural language, concepts are expressed through definitions, which specify a unique intension and extension. A designation (term, appellation or symbol) represents a concept. A subject field is a branch of human knowledge and is comprised of a set of related concepts or concept system. A set of designations makes up a special language, which is used in a subject field. A terminology is the set of designations that represent the concepts in a concept system. A terminology system is the combination of a concept system and a terminology that represents it.

3.2.3 Evaluation of ISO/IEC 11179 family of standards

As illustrated in section 2.2.3, metadata is important for the longevity of digital objects. A wide range of metadata schemas are potentially relevant for the longevity of digital objects. The application of metadata schemas for the creation of preservation metadata is hampered by the fact that they are often ambiguously formulated and defined. We have seen that memory institutes use a wide range of metadata elements to describe object classes, such as photographs and digital surrogates, and that these metadata elements are often not formulated and applied in a clear and unambiguous way.[94] A reason for this is that the things, ideas and abstractions that belong to the object class have vague boundaries. In order to 'discriminate exactly what we know vaguely' ([SOW00] p349),[95] the formulation of metadata elements according to the ISO/IEC 11179 family of standards could be a valuable building block to improve the quality of the metadata schemas and thus the durability of the object classes. The descriptors that are part of metadata schemas can be considered as Data Elements, as defined by the ISO/IEC 11179 family of standards.

Based on an analysis of initiatives in the field of standardisation, the ISO/IEC 11179 family of standards seems an appropriate source for the formulation, definition and expression of Data Elements in an unambiguous way. The purpose of ISO/IEC 11179 is to describe the standardisation and registering of Data Elements to make data understandable and shareable by machines and people. Despite the fact that the ISO/IEC 11179 family of standards is currently under revision, it is relevant for the structure of metadata to study the basic attributes of Data Elements. The establishment and revision of ISO/IEC11179 illustrate the strong and weak aspects of a formal standardisation process.

Part 3 of the standard, concerning basic attributes of Data Elements, is the most relevant for this study as it can be used to formulate durable Data Elements that enable the longevity of objects that are documented with these Data Elements. The ISO/IEC 11179-3 standard provides a framework for the accurate definition of Data Elements. It can be used for the creation of building blocks, the aim of which is to create and maintain metadata schemas for digital visual sources. Later in this study the concepts of this standard are applied in case studies and combined with other insights from the preservation metadata community.

[94] Table 2.2 contains a number of examples of metadata schemas that are used by memory institutes.

[95] Sowa is quoting philosopher Alfred North Whitehead.

3.3 Durable access to digital objects

In this section the premise 'Permanent access to digital objects can be realised by persistent identifiers, preservation repositories and open access' is examined. An obvious durability feature of digital objects concerns the ability to have access to the digital objects and the ability to use them, even in the long term. Preservation without access seems pointless. The aim of this section is to discuss building blocks that facilitate the long-term access and usage of digital objects via the Internet. Two components for durable access to digital objects are discussed. First, persistent identifiers for digital objects are explained. Next, the relevance of preservation repositories is covered. The two components enable durable, long-term access to digital objects such as digital surrogates of historical photographs.

The way in which long-term access is provided to objects, whether the objects have an analogue or digital nature, very much depends on the archival principles applied by the organisations that have the objects in their collection. Archival arrangement has a long tradition, based on a number of principles, and this is very well described in *Arranging and describing archives and manuscripts* by Miller [MIL90]. Preservation of electronic records is the main focus of the InterPARES project.[96] Electronic records are records created and/or maintained in databases and document management systems in the course of administrative activities. InterPARES1 addressed the development of theoretical and methodological knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form. InterPARES2 focuses on the reliable creation, as well as on the authentic preservation, of records in dynamic and interactive systems, in the course of artistic, scientific and e-government activities ([DUR04], p. 218).

Fundamentals for the appraisal of photographic collections, an important phase that determines the way access is provided to photographic collections, can be found in [JON96]. One of the few Dutch publications dedicated to the description of photographic collections is [HOG94]. As access to analogue photographic collections from an archival perspective is covered by existing studies such as those mentioned above, this section focuses on durable access to digital surrogates of historical photographs.

The two durable access components – persistent identifiers and preservation repositories – discussed later in this section are closely related to the 'life-cycle management' and 'records continuum' concepts, as established by records management for many years. As stated by Jones and Beagrie, life-cycle management implies '... the need actively to manage the resource at each stage of its life-cycle and to recognise the inter-dependencies between each stage and commence preservation activities as early as practicable. This represents a major difference with most traditional preservation, where management is largely passive until detailed

[96] InterPARES 1 (see: FIN02]) was launched in 1999 and InterPARES 2, which began in 2002, is expected to be completed in 2006. The website of InterPARES can be found at: <http://www.interpares.org> [cited 24 November 2004].

conservation work is required' ([JON01] p. 11). The aim of the records continuum concept is to make connections between business explicit, defined broadly to encompass all social and organisational activity, the people or agents who do business, and the records that are by-products of that business. Its vision links the dynamic world of business and social activity to the passive world of information resource management in cyberspace [MCK99].

A high-quality digital surrogate of an historical photograph can be characterised as a 'use-neutral' digital master image and this implies a long life-cycle with appropriate life-cycle management. The naming of the digital surrogates using persistent identifiers and the storage of the objects in preservation repositories can be considered as important components of a system that gives durable access to digital objects. These components must be identified and applied as early as possible in the life-cycle of a durable digital object.

3.3.1 Persistent identifiers

[PER04] covers the establishment of methods for persistent identification of government resources. Identifiers establish the identity of objects by attaching labels to them. Persistent identification of digital objects is an important component of a digital infrastructure to ensure its effective long-term storage and access. Persistent identifiers can help to remove boundaries between and among communities and disciplines. Requirements for the persistence of the identifiers are their authority, reliability and functionality throughout the life-cycle of the object. A persistent identifier tracks a specific object regardless of its physical location or current ownership. The decision on what level of persistence of access to provide in the long term will be a matter of judgement, and this should be determined at the point of creation of the object.

The importance of persistent identification of objects is illustrated by part 5 of the ISO/IEC 11179 family of standards [ISO11179-5:1995]. This standard provides rules and guidelines for naming and identification of Data Elements. Names are established by the use of a naming convention. The names of the Data Elements must be unique within the register of an organisation that develops and maintains that register. The purpose of using a naming convention is name consistency, from which users can infer facts about the definition, usage and relationships of the items. The data identifier, defined as 'an identifier of a Data Element (a string of characters or other graphic symbols) assigned by a Registration Authority', uniquely identifies a Data Element ([ISO11179-5:1995] p. 4). The ISO/IEC 11179 family of standards does not specify the format or content of a unique data identifier. A number of persistent identifier systems are available that are more specific than part 5 of the ISO/IEC 11179 family of standards.

The application of persistent identifiers for digital surrogates of historical photographs consists of five steps (ref. [PAY98]):

1. *Selection of a persistent identifier scheme*. The most important systems that

can be used to create persistent identifiers are described below. The cost and benefit of a system must be evaluated.

2. *Establishment of a naming authority*. Rules and policies for the naming of the digital surrogates must be established within the selected system.
3. *Creation of persistent identifiers*. Identifiers must be assigned to the digital surrogates according to the identifier scheme chosen.
4. *Registration of the persistent identifiers*. The identifiers must be translated to locations. For this, a resolution service is required, either local or global.
5. *Usage of the persistent identifiers*. Apart from the moving of servers, changes in applications and other modifications of the infrastructure, the persistent identifier is stable and able to locate a specific digital surrogate.

Six systems that can be used for the persistent identification of digital objects are the 'Universal Resource Name' (URN), the 'info' URI, the 'Persistent Universal Resource Locator' (PURL), the Handle System, the 'Digital Object Identifier' (DOI) and the 'Archival Resource Key' (ARK).

URN: Uniform Resource Name

The URN specification is related to the 'Uniform Resource Locator' (URL)[97] specification that was developed in 1994 as the first protocol to locate and access resources on the Web. The URL protocol does not cater for changes in the names or storage locations of resources and this is the main cause of broken hypertext links on the World Wide Web. Broken hypertext links illustrate the relevance of persistent addressing and identification of web documents. Standardisation of Internet-related specifications is published as part of the 'Request for comments' (RFC) document series. This series is the official publication channel for Internet standards documents and is managed by the Internet Engineering Task Force (IETF), the standard-setting body for Internet development.

In 1998, the general syntax of the 'Uniform Resource Identifier' (URI) was published as RFC2396.[98] The URI specification aims at identifying an abstract or physical source. The URI scheme identifies the type of resource and access method.[99] The URN is one of the registered URI schemes[100] aimed at persistent naming of resources. The IETF defines a URN as a 'globally unique, persistent identifier used for recognition of, or access to, a resource or a unit of information'.

[97] The specification of the URL protocol can be found at: <http://www.ietf.org/rfc/rfc1738.txt> [cited 19 July 2004].

[98] The specification of the URI protocol can be found at: <http://www.ietf.org/rfc/rfc2396.txt>. The URI protocol revises and replaces RFC1738 [cited 19 July 2004].

[99] There are more than 40 URI schemes available, see <http://www.iana.org/assignments/uri-schemes> [cited 19 July 2004]. Some of the most important schemes are 'http', 'ftp' and 'news'. 'urn' is also one of the URI schemes.

[100] The specification of the URN protocol can be found as RFC2141 at: <http://www.ietf.org/rfc/rfc2141.txt> [cited 19 July 2004].

A URN is structured as:

<URN>::=‘urn’:<NID>‘:’<NSS> [101]

<NID> is the namespace identifier and <NSS> is the namespace specific string. So-called ‘national bibliographic numbers’ (NBNs) are used as URNs to create persistent identifiers for digital objects.[102] National bibliographic numbers are persistent and unique identifiers assigned by National Libraries. The application of URNs as persistent identifiers requires a registering authority that manages the assignment and resolving of the URNs. The Library of Congress is the designated registration agency of the NBN namespace (urn:nbn). Sub-identifiers are assigned at county level by a national library. This sub-identifier is expressed as an ISO country code, for instance ‘de’ for Germany. In Germany, the library community uses the NBN implementation of the URN specification for the creation and management of persistent identifiers.[103]Both URLs and URNs are types of URIs. While URLs are locators, or addresses, on the Web, URNs are names on the Web.

‘info’ URI

The URN is one of the registered URI schemes. There exist many identifier schemas that are not available as URI schemes, such as the Dewey Decimal Classification,[104] and for this the ‘info’ URI scheme has been developed.[105] Instead of using the URI registration mechanism to express an identification system, a new construct has been developed to facilitate the referencing of objects that have identifiers in public namespaces, by means of their ‘info’ URI. The ‘info’ URI scheme enables public namespaces that are not part of the URI allocation to be represented and provides a bridging mechanism to allow public namespaces to become part of the URI allocation. The namespaces declared under the ‘info’ URI scheme are regulated by an ‘info’ registration mechanism. The ‘info’ URI scheme is the basis for the naming architecture of the OpenURL system (ref. [SOM01]). OpenURL is a generic web technology that enables the context of a service request to be evaluated within the user environment. The ‘info’ URI scheme is a special type of URN that complements regular URNs but is designed to be simpler and more convenient both to manage and to use.

[101] Phrases enclosed in brackets are required. An example of a URN is: ‘urn:nbn:de:1111-20040330226’.

[102] The usage of NBNs as URIs is published as RFC3181. See: <http://www.ietf.org/rfc/rfc3188.txt> [cited 19 July 2004].

[103] See: <http://www.persistent-identifier.de> [cited 2 September 2004].

[104] The Dewey Decimal Classification (DDC) system is a widely used library classification system, see: <http://www.oclc.org/dewey/> [cited 12 December 2004].

[105] The ‘info’ URI Scheme for Information Assets with Identifiers in Public Namespaces. Internet Draft, see: <http://www.ietf.org/internet-drafts/draft-vandesompel-info-uri-03.txt> [cited 10 January 2005].

NISO (National Information Standards Organisation) will act as the maintenance agency for the 'info' registry.[106] The 'info' URI scheme is less persistent than URN namespaces. URNs are intended as persistent, location-independent resource identifiers, whereas the 'info' URI scheme does not assert the persistence of the identifiers. It is, rather, a lightweight registration mechanism for public namespaces of object identifiers. The 'info' URI scheme is neutral with respect to identifier persistence.

PURL: Persistent Universal Resource Locator[107]

The PURL system was developed by the computer library service and research organisation OCLC[108] as a contribution to the IETF URN activity. Since 1996, PURL has been available as a tool for managing names and namespaces. A PURL looks like a URL whose server address is the name of a so-called resolver service. A PURL is structured as:

http://purl.[resolver name]/[specific resource identifier]

The resolver server performs a database search to ascertain the location of the identified object. If the URL of an object changes, the only maintenance required would be an update to the PURL location in the database of the resolver server. By making URLs symbolic names it will improve their usefulness. The PURL toolset takes advantage of the redirection facility in the HTTP protocol. In essence, a PURL server is a redirection server. By the middle of 2004, about 600,000 PURLs had been created.

PURLs do not improve the persistence of the objects identified by them, but rather the names of the resources. The PURL toolset can be used to manage resource names and resource locations with greater reliability. The uniqueness of the identifiers is determined by the characters of the URLs. The registrant of the PURL can change the 'identifier–object' bindings. Thus, PURLs do not replace policies for managing resource names.

[106] The 'info' URI registry can be found at: <http://info-uri.info>. The registry provides a mechanism for the registration of public namespaces that are used for the identification of objects, and that are not part of the URI allocation. An example of an 'info' URI is: 'info:ddc/22/eng//004.678'. 'ddc' is the Dewey Decimal Classification namespace and '22/eng//004.678' identifies an object within that namespace.

[107] See: <http://www.purl.org> [cited 23 September 2004].

[108] See: Online Computer Library Center, OCLC: <http://www.oclc.org> [cited 14 December 2004].

Handle System[109]

The Handle System was developed by the 'Corporation for National Research Initiatives' (CNRI).[110] It is a framework for managing digital information and provides a naming scheme for unique identifiers, called 'handles'. A resolution system translates the handles into location-related data. A centrally administered registry service manages the resolving naming authorities. A Handle consists of two parts: a naming authority and a unique string that identifies an object. A Handle is structured as:

> [unique persistent naming authority for the assigning agency]/[unique persistent identifier for the resource][111]

The Handle System is a distributed, scalable and secure identifier system. It enforces the use of unique names for objects and is based on an open protocol.[112] The Handle System may be implemented by anyone who agrees to the basic licensing terms, thus only providing persistence if used with an appropriate social infrastructure. The Handle System is used by a large number of organisations, such as the Library of Congress, the International DOI Foundation and the DSpace platform.

DOI: Digital Object Identifier

The DOI can be considered as a community-related implementation of persistent identifiers. It is an initiative of the 'Association of American Publishers' (AAP) and is governed by the 'International DOI Foundation' (IDF).[113] The Handle System is used as its underlying technology for the administration and resolution of the DOIs. The DOI System uses the Handle System as one of the components in building an added-value application for the persistent, semantically interoperable identification of objects. These components are a numbering syntax, a resolution service, a data model, and policies and procedures for the implementation of DOIs through a federation of registries. The IDF requires a consistent DOI prefix and numbering syntax. The IDF has a license for use of the Handle System, together with the ability to sub-license this for DOI assignment to all DOI Registration Agencies. Use of the DOI system therefore does not require a separate Handle System license.

[109] See: <http://www.handle.net> [cited 23 September 2004].

[110] See: <http://www.cnri.reston.va.us/> [cited 23 September 2004].

[111] An example of a valid handle is: 'archive.images/dm1394939.tif', where 'archive.images' is the naming authority and 'dm1394939.tif' the string that identifies a particular object.

[112] RFC3650 'Handle System Overview', see: <http://www.ietf.org/rfc/rfc3650.txt> [cited 23 September 2004].

[113] See: <http://www.doi.org> [cited 23 September 2004].

ARK: Archival Resource Key[114]

The extent to which structured digital data remains predictably available through known channels is a central concern for most organisations whose mission includes an archival function ([KUN01] p. 1). The ARK, developed by the California Digital Library (CDL),[115] is targeted at the persistent identification of archived digital objects. An ARK identifier contains a link to the digital object metadata, the digital object content files and a commitment statement concerning the digital object. This commitment statement is a promise regarding the degree of persistent maintenance. The ARK is used by the CDL to manage persistent access to its digital objects, including digital images.

An ARK is represented by a sequence of characters that contain the label 'ark:', optionally preceded by the beginning part of a URL. The URL, or the 'Name Mapping Authority Hostport' (NMAH), is variable and replaceable. The invariable, globally unique identifier follows the 'ark:' label. This includes the 'Name Assigning Authority Number' (NAAN) followed by the name of the object. The ARK syntax can be summarised as:

[http://NMAH/]ark:/NAAN/Name

The NMAH part is in brackets to indicate that it is optional and replaceable. The name assigning authority might be a national library, a national archive or a publisher or repository. It establishes long-term associations between identifiers and objects. The name assigning authorities are designated by Name Assigning Authority Numbers, and Name Mapping Authorities can resolve the ARK. The redirection is realised by a simple search table, but a mechanism similar to the Domain Name System (DNS) is planned.

Persistent Identifier approaches

All the discussed digital identifier policies and all systems for the creation and application of persistent identifiers require registration and resolution services. Successful implementation requires institutional support and management. The development of a persistence policy can be organisationally time consuming. It seems that persistent identification of digital objects is a matter of service intent of the 'Name Mapping Authorities'. The Handle system, with a global identifier resolving mechanism, and the URN system, with an identifier resolving mechanism at community level, are currently the two main systems that can be used to create persistent identifiers for digital surrogates of historical photographs.

In principle, the only guarantee of the usefulness and persistence of identifier systems is the commitment of the organisations that assign, manage and resolve identifiers.

[114] See: < http://www.cdlib.org/inside/diglib/ark/> [cited 25 September 2004].

[115] See: <http://www.cdlib.org> [cited 12 December 2004].

3.3.2 Preservation repositories

Increasingly, the 'repository' concept is used in relation to persistent access to digital objects. This section elaborates on the current state of the art regardingg digital preservation and persistent access to digital objects, such as digital surrogates of historical photographs, stored in repositories. The close relationship between preservation and access is very well illustrated by Hedstrom: 'In the past, archivists argued that there was no point preserving records if one could not provide access; in the digital environment, without access there will be no preservation'. ([HED95] p. 189), quoted in ([CHI01] p. 146). Hodge and Frangakis provide an overview of systems and projects addressing preservation and permanent access in science and technology. Institutional repositories play an important role in the long-term access to digital data [HOD04].

Digital conversion is an expensive activity, especially if a high-quality benchmarked conversion method is applied and rich preservation metadata is created. Permanent access to digital objects is increasingly associated with the term 'digital preservation'. Permanent access implies adequate rendering of the digital object, given the technological changes that have occurred and will continue to occur.

According to Lynch, an institutional repository is a 'set of services that the institution offers to members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organisational commitment to the stewardship of these digital materials, including long-term preservation where appropriate. ... At any given point in time, an institutional repository will be supported by a set of information technologies, but a key part of the services that comprise an institutional repository is the management of technological changes, and the migration or digital content from one set of technologies to the next as part of the organisational commitment to providing repository services' [LYN03]. Here, migration is mentioned as the strategy for keeping digital objects accessible. Preservation is mentioned as a function of an institutional repository. In a sense, institutional repositories have already existed for decades as data archives (ref. [DOO04]).

Smith states '... the goals of preservation have changed. They are no longer to fix, to stabilize, to conserve, or to reformat onto an archival medium. We speak now of ensuring continued access through maintaining collections that are fit for use. Increasingly, we hear that our users and potential users want more access to more resources, and they want them delivered in ways that promote customisation and re-purposing' ([SMI04] p. 7). The decision regarding what level of persistence of access to provide to an object will be a matter of judgement and this should be determined at the time of creation of the object.

Open Access

The relation between preservation of digital objects and the design criteria for preservation repositories is emphasised by Dondorp and Van der Meer [DON03].

They mention 'technology independence through standardisation' as an important design criterion for preservation repositories. Important standardisation issues of preservation repositories concern the protocols and architecture of the infrastructure that provides access to distributed repositories. Storage of digital objects in distributed, interoperable repositories is gaining more and more ground as the most efficient model to provide permanent access to digital objects where creators have control over the integrity of their work and the right to be properly acknowledged. In the scientific community, this 'open access' model has been implemented in several cases. A protocol for metadata harvesting, developed by the Open Archives Initiative (OAI-PMH), is an important construct for the implementation of permanent access and thus durable storage, as well as the digital preservation strategy known as 'LOCKSS', the acronym for 'Lots Of Copies Keep Stuff Safe'. Both architectures are discussed.

In [SOM04] the transfer of digital objects to one or more trusted digital repositories that store and preserve safety copies is mentioned as an important use case for the need to develop methods for indirectly gathering digital resources through the metadata that is exposed by the preservation repositories.

Open Archives Initiative Protocol for Metadata Harvesting[116]

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives programme. The fundamental technological framework and standards that have been developed to support this work are, however, independent of both the type of content offered and the economic mechanisms surrounding that content, and promise to have much broader relevance in opening up access to a range of digital materials.

The goal of the 'Open Archives Initiative – Protocol for Metadata Harvesting' (OAI-PMH) is to supply and promote an application-independent interoperability framework that can be used by a variety of communities engaged in publishing content on the Web. The OAI-PMH protocol permits metadata harvesting. Increasingly, organisations are developing dissemination systems based on this protocol, as they are aspiring to unlock access to their collections in a simple and low-maintenance manner, and to make interchange of metadata between multiple heterogeneous archives possible.

The OAI-PMH is a communication protocol or language with only six permitted verbs. The verbs 'Identify' and 'ListMetaDataFormats' and 'ListSets' are related to repositories that are part of a Data Provider. The other three verbs are the actual harvesting verbs: they make it possible to locate an individual resource within a

[116] The website of the Open Archives Initiative can be found at: <www.openarchives.org> [cited 12 September 2004].

repository. When the Service Provider contacts one of the Data Providers for the first time, it will use the 'Identify' verb. The Data Provider will respond with a data block containing relevant information about the organisation hosting the Data Provider and the Data Provider itself.

After the Service Provider has received and interpreted the identification information, it will ask the Data Provider which metadata formats it supports by issuing a 'ListMetadataFormats' verb. This yields a list of metadata formats, their validation scheme locations and a metadata prefix. This is a unique code that identifies the metadata format in the repository; for instance, whenever a Service Provider wants information in a specific format, it should specify this format using the appropriate metadata prefix.

The OAI-PMH protocol effectively removes the dependencies on system architecture and metadata compatibility. Archives opening up their collections using OAI-PMH merely have to install a Data Provider module acting as a uniform access point to the metadata. A Data Provider maintains one or more repositories that support the OAI-PMH as a means of presenting metadata. The Data Provider is a passive software component connecting the metadata repository to the Internet by using the XML data format and the Hypertext Transmission Protocol (HTTP),[117] and can be implemented to offer open access using multiple metadata formats. Data Providers are systems that support the OAI-PMH as a means of presenting metadata.

As soon as the Data Provider is realised, it can be accessed and harvested by Service Providers. A Service Provider issues OAI-PMH requests to Data Providers and uses the metadata as a basis for building value-added services. This means that another system can connect to the Data Provider and download the metadata it needs in one of the available metadata formats. The downloaded metadata is stored and indexed by a Service Provider that will offer the actual end-user access to the metadata. One Service Provider can harvest as many collections as it sees and offer access to all of these repositories at once, making it seem as if the end-user is accessing one large collection.

The use of the OAI-PMH for providing archival access to metadata that is encoded in the EAD (Encoded Archival Description)[118] format is described in [PRO03]. This article illustrates the importance of the OAI-PMH for open access to preservation repositories.

Lots of Copies Keep Stuff Safe (LOCKSS)

A digital preservation strategy developed by Stanford University bears the acronym LOCKSS [REI00].[119] The system is aimed at providing sustainable access to

[117] For more information on HTTP, see: <http://www.w3c.org/Protocols/> [cited 12 July 2004].

[118] The Encode Archival Description (EAD) structure is based on ISAD (see: Table 2-2).

[119] The website of the Lockss system is: <http://www.lockss.org> [cited 20 May 2004].

scientific journals on the Web, but in principle can also be used in a digital preservation strategy for other digital objects, such as digital images. The developers of the LOCKSS system realise that digital longevity does not have a single solution and diversity is essential to successful preservation. The LOCKSS system is based on librarians' techniques for preserving access to published papers. It includes: acquire lots of copies, distribute them around the world so that it is easy to find some of the copies, but hard to find all of the copies, and lend or replicate your copies when other libraries need access.

LOCKSS creates low-cost, persistent digital 'caches' of authoritative versions of http-delivered content. A distinction is made between preserving archives and preserving general circulating collections. This second type of preservation is the domain of the LOCKSS system. The LOCKSS system is realised as Open Source software in order to improve confidence in the system. It is assumed that a proprietary system will not be accepted. Moreover, programmers are able to refine and improve the system. The LOCKSS software offers a cheap and easy way of running web caches.

The LOCKSS system has three components:

- *Collecting of content into cache*. Web crawlers coming from authorised IP addresses gather all new content for an instance of a LOCKSS system. Off-the-shelf technology is available to implement this component.
- *Serving content to users*. A web server exports the contents of each cache to the network's users.
- *Preserving the cache for posterity*. The heart of LOCKSS is a peer-to-peer inter-cache protocol called LCAP (Library Cache Auditing Protocol). It allows the caches to agree on which URLs should exist and what their contents should be. If a cache discovers a missing or damaged URL it can fetch a new copy via http from the original publisher or from one of the other caches.

The beta testing phase of the LOCKSS software has been completed successfully. However, the LOCKSS system is not yet ready for general use. Modifications to the software to improve performance, security and usability are still required.

3.3.3 Evaluation

A number of building blocks are available to create persistent identifiers for digital surrogates as well as to create preservation repositories in which digital surrogates can be stored for the long term. These building blocks were presented in the preceding sections. Neither persistent identifiers or repositories are developed with the specific primary purpose of facilitating the durability of digital objects, but they fit very well into an infrastructure for long-term storage of digital objects.

The URN and the Handle system are the most relevant for the creation of persistent identifiers of digital surrogates of historical photographs. The URN is relevant because it is related to a highly formalised and robust registration mechanism that maximises the commitment of the organisation considering itself responsible for the assignment of persistent identifiers to digital objects. The Handle system is a fairly complete architecture ready to be used for the persistent identification of digital objects. The commitment of the organisations that assign, manage and resolve persistent identifiers – irrespective of the identifier system used – in all application domains is the key to successful implementation of a persistent identifier policy for digital surrogates of historical photographs.

Preservation repositories are important for long-term access to digital objects. The function and attributes of preservation repositories are becoming increasingly clear, as is illustrated by the literature presented earlier in this section. The open access paradigm implies easy access to repositories of digital objects in a distributed environment, using open standards taking account of the integrity of the objects as well as property rights. Lowering the threshold for access to preservation repositories contributes to the usage of the digital objects and thus to their longevity. The OAI-PMH protocol and LOCKSS system are important building blocks for the implementation of preservation repositories of digital surrogates of historical photographs.

4 Experiments on the longevity of digital surrogates of historical photographs

Chapter 2 described why and how memory institutes digitise historical photographs. Also, the state of the art regarding the longevity of digital objects was described. Chapter 3 introduced and clarified a number of premises upon which activities can be based that will result in durable digital surrogates of historical photographs, or ‘permanent pixels’. These premises concern the benchmarking of the digital capture process, the unambiguous formulation of preservation metadata and persistent access to digital objects based on open standards.

In this chapter, a number of approaches for realising the creation of durable digital surrogates of historical photographs are evaluated in the form of ‘experiments’. The experiments are intended to examine the validity of hypotheses upon which the durability of digital objects can be based and to demonstrate how the method can be used in practice. As technology will continue to evolve in the future, the problem of digital preservation cannot be considered as a static field. Therefore, the experiments must be seen in the light of evolutionary change. Also, in the future, just as today, people will want to use the best technology available, so there will be new hardware, new software and new data formats. This evolution will have its roots in the current state of technology and gradually evolve from this.

The experiments described in this chapter illustrate to what extent durable digital objects can be created, and should be considered as independent case studies. An evaluation of the experiments can be found in chapter 5. The goal of the experiments is to assess a number of methods for creating durable digital objects. For these experiments a number of existing building blocks are used. The experiments are based on the assumption that existing building blocks are potentially more durable than a specific methodology developed in isolation. Building blocks are tools, procedures, specifications and guidelines available to perform the experiments.

The experiments described in this chapter are:

- *Experiment 1: The durable digital image file format*. This experiment is based on the hypothesis that a standard image file format is a durable file format.

- *Experiment 2: The durable bitstream of the digital image.* This experiment is based on the hypothesis that the bits representing the individual pixels of a digital image can be expressed in the standardised mark-up language XML. The XML data format is considered to be durable.
- *Experiment 3: The durable formulation of preservation metadata.* This experiment is based on the hypothesis that unambiguously formulated metadata elements of digital objects are essential to assess, understand and process these digital objects in the future. Preservation metadata consists of all metadata elements relevant for the durability of a specific digital object. Metadata elements that are used and reused by a wide user community can be considered as durable.

Experiments 1 and 2 address the durability of the data format that contains the pixels making up the digital surrogate. Experiment 3 takes the context of the digital image in a given data format into consideration. This context provides the durability of the digital image.

4.1 Experiment 1: Durable digital image file format

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 broadly supported object. The requirements for standard data formats are given, followed by an assessment of existing data formats for digital raster images.

4.1.1 Requirements for standard digital image file format

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.
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. A wide range of image capture devices as well as image processing systems, for instance, 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 relevant for this study, namely digital raster images, three more 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 surrogates of historical photographs:

4. The data file must be uncompressed. Data compression is strongly advised

against for two reasons. Data compression can lead to a loss of image quality and a compressed digital image has a greater risk of becoming unreadable than an uncompressed digital image. Both issues are clarified below.

Raster images tend to be very large, so data compression algorithms are applied in order to reduce the required data storage. Most data compression algorithms for raster images are based on the principle that the human eye cannot discriminate between all the 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 likewise the file size. In almost all cases, compression of digital images of photographs will lead to a 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.[120]

A corrupt bit in a compressed image file results in a 'dead image', whereas the chances are that a corrupt bit in an uncompressed image will just be a 'dead pixel'. Thus, an uncompressed raster image is considered more durable than a compressed raster image, because the former has a greater chance of being interpreted if some bits are altered in the course of time.

5. A durable data format for digital surrogates of historical photographs must contain facilities for storing 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 all significant characteristics of the analogue original that it is based on. This means that all colours, details and the dynamic range of the original can be represented in the digital image. Section 3.1 of this dissertation contains a discussion of the benchmarked digital capture process.

A considerable number of graphic file formats have been developed in the past and in the future new formats will also 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 the file format requirements for the creation of high-quality digital surrogates of historical photographs is that the file format must exist for a considerable period of time. The list of file formats given in the *Encyclopedia of Graphics File Formats* [MUR94],[121] published about ten years ago, is used as a reference for raster image file format standards that are potentially relevant, as the book contains the description of all common image data formats.

[120] The website of the JPEG2000 standard can be found at: <http://www.jpeg.org/JPEG2000/index.html> [cited 19 March 2004].

[121] Both the first and second editions of the Encyclopedia are used. The first edition was published in 1994, the second in 1996. Until now an update has not been published.

Table 4.1 Durability specifications of raster file formats

	Raster file requirements	TIFF	JPEG	GIF	PNG
1	Used by a large community over a long period of time	X	X	X	-
2	File format specification is published	X	X	X	X
3	Supported by a wide range of applications	X	X	X	X
4	Supports uncompressed / single-page images	X	-	-	-
5	Facilities for preservation metadata	X	-	-	X
6	Enables 'full informational capture'	X	-	-	X

Four raster file formats mentioned in the encyclopaedia that are still used today are: TIFF, JPG, GIF and PNG .[122] This observation is supported by a number of publications such as [GUI00A] in which the TIFF format is mentioned as the most appropriate for digital master images. JPEG has a very large user community, as all standard web browsers support it. Web browsers also support GIF, but this file format uses a formerly patented compression algorithm and is less widely used on the Web. The PNG format seems to be very appropriate to serve as a file format standard for durable digital images, mainly because an independent group designed it. However, by default PNG applies a (loss-less) compression algorithm and the PNG user community is not very large. To what extent the four file formats meet the durability requirements is stated in Table 4.1.

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.

4.1.2 The TIFF image file format

The raster image file format TIFF ('tagged image file format') meets all requirements for durable digital objects, as mentioned in Table 4.1. The TIFF file format is specified in [TIF92]. The file format is largely hardware and software independent and has been around for more than ten 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 of creating 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 Systems Inc.[123] TIFF version 6 was made publicly available in 1992. The original TIFF specification was released in 1986 by Aldus Corporation (later acquired by Adobe) as a standard method for storing black-and-white images created by scanners and desktop publishing standards. The functionality of subsequent

[122] Graphics file formats originally developed for a specific platform, 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 initial selection as well as non-raster image file formats.

[123] The specifications of the TIFF file format version 6.0 can be found at: <http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf> [cited 14 January 2004].

versions of the TIFF raster image file format has improved considerably. 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 in graphics newsgroups archives on the Internet reveal that Adobe was not happy to inherit the TIFF standard when it acquired Aldus Corporation, because TIFF is a competitor of the PDF standard developed by Adobe. However, Adobe cannot abandon support of the TIFF standard as it has a very large user community.

The success of the TIFF specification as a widely used standard for digital raster images is attributed to its extensible nature and support for numerous data compression schemes. As a consequence, developers are able to customise the format to fit any particular data format need. Both compressed and uncompressed images can be coded in the TIFF standard and it is also possible for more than one image to be 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, for instance locally used metadata. This means that features that restrain the longevity of digital images as stated above are also part of the standardised TIFF 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 image file format (for instance, JPEG) is needed before a web browser can process an image. A TIFF image 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 lower 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 a carefully chosen format, also called magic number, which 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 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 4.2 in ascending order by decimal code.

Table 4.2 Assessment of Baseline TIFF 6.0 information fields

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'		
SamplesPerPixel (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

For the digitisation of photographs, 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 characteristics of the original. Section 2.1.3 gives a classification of photographs upon which the type of TIFF file should be based. Baseline TIFF sets a number of required fields for these image types that can be found in the third column of Table 4.2.

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 fairly complex and three mandatory 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 (see criterion 4 in section 4.1.1), this information field must have the value '1', meaning 'No compression'. The remaining mandatory 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 4.2 indicates whether the data in the information field can be considered as preservation metadata for digital surrogates of historical photographs, 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. In general, the more documentation the better, but information fields that are not relevant for the image type or information fields that hamper the digital durability, for instance 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 4.2 gives 24 information fields of baseline TIFF that support digital durability. The application of five of the information fields actually hampers digital durability and seven of the information fields are not relevant for digital durability.

In principle, the baseline TIFF version 6.0 file format meets all six requirements for durable 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 significant characteristics of the original.

TIFF extensions

TIFF extensions are features of the TIFF image file format that may not be supported by all TIFF readers. The official TIFF 6.0 standard contains a number of TIFF extensions, and separately a number of TIFF extensions are published independently of the official TIFF 6.0 standard. The officially published TIFF 6.0 specification contains four groups of extensions. One group of extensions relates to 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 to the TIFF 6.0 specification, has excellent applicability for device-independent manipulation of continuous tone images. For digital surrogates of colour photographs, this high-quality colour space is of great importance.[124] If the CIELAB colour space is used, the information tag 262 (Photometric Interpretation) should have the value '8'. This value has the meaning '1976 CIE LAB'.

In addition to extensions that are part of the official TIFF standard there are also a number of separately published extensions. To mention some examples: a specification of TIFF especially for GIS applications, a TIFF specification that supports the JPEG compression method, and a TIFF extension as a pre-press interchange format.

In principle, the TIFF specification meets the criteria to act as the standard for the creation of durable digital surrogates, but features that obstruct durability, such as data compression and inclusion of non-standard private tags, should not be used. This means that the TIFF standard is fairly tolerant when it comes to compliance to the standard.[125] This makes the format flexible and therefore adaptable in variable situations, but it may also lead to sloppiness.

One extension to the TIFF 6.0 format that has the status of an international ISO standard and can be considered as a durable file format is the TIFF/EP image data format ([ISO12234-2:2001]). TIFF/EP was 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.

TIFF version 6.0 as durable image file format

Despite the fact that on the Web an enormous number of digital photograph collections are available, it is difficult to get online access to the archival images on which the derivatives are based. One of the few institutes that provide web access to digital archival images is the American Memory project of the Library of Congress. Their archival images in TIFF format can be downloaded and analysed. The website of the American Memory project contains a lot of detailed background information on the way the conversion of the original photographs was carried out, and this information plays a role in assessing the value of the digital master files. The conversion approach always takes the appearance and value of the original photograph into consideration.

[124] Section 3.1.3 discusses the importance of the device-independent colour space CIELAB.

[125] The PNG raster image file format (Portable Network Graphics), a W3C recommendation, is considered to be a well-designed, easily accessible, superior format, but after a promising outset, this format did not gain a wide user community. See: <http://www.w3c.org/Graphics/PNG> [cited 14 February 2004].

```
00208a.tiff:
Magic: 0x4949 <little-endian> Version: 0x2a
Directory 0: offset 132610346 (0x7e7792a) next 0 (0) ImageWidth (256) SHORT (3) 1<9681>
ImageLength (257) SHORT (3) 1<6849>
BitsPerSample (258) SHORT (3) 1<16>
Compression (259) SHORT (3) 1<1>
Photometric (262) SHORT (3) 1<1>
DocumentName (269) ASCII (2) 17<ppprs/00208a.tif\0>
StripOffsets (273) LONG (4) 2283<8 58094 116180 174266 232352 290438 348524 406610 464696
522782 580868 638954 697040 755126 813212 871298 929384 987470 1045556 1103642 1161728
1219814 1277900 1335986 ...>
Orientation (274) SHORT (3) 1<1>
SamplesPerPixel (277) SHORT (3) 1<1>
RowsPerStrip (278) SHORT (3) 1<3>
StripByteCounts (279) LONG (4) 2283<58086 58086 58086 58086 58086 58086 58086 58086 58086
58086 58086 58086 58086 58086 58086 58086 58086 58086 58086 58086 58086 58086 58086
58086 ...>
XResolution (282) RATIONAL (5) 1<1425>
YResolution (283) RATIONAL (5) 1<1425>
PlanarConfig (284) SHORT (3) 1<1>
ResolutionUnit (296) SHORT (3) 1<2>
DateTime (306) ASCII (2) 20<2000:03:21 21:36:29\0>
Artist (315) ASCII (2) 20<Library of Congress\0>
```

Figure 4.1 TIFF information tags of a digital master file of the American Memory project

One randomly chosen collection is analysed in detail. This collection is considered representative of all collections available in the American Memory project. The collection of photographs on Japanese-American internment in 1943 compiled by the famous photographer Ansel Adams consists of both negatives and prints, and archival images have been created of all the photographs in both formats.[126] Figure 4.1 contains the TIFF information tags of a digital master file of the American Memory project.[127]

The digital master images of the American Memory collection are not compressed, have a dynamic range of 16 bits per pixel and have a spatial resolution of about 10,000 pixels on the long side, with the short side scaled in proportion. This results in images of about 130 megabytes.[128] Each file contains one photograph and no image enhancement has been applied. All obligatory information fields of the baseline TIFF 6.0 specification are present in the digital master files of the American Memory project. Of the baseline tags that contain preservation metadata only 'DateTime' and 'Artist' are used. One TIFF information tag from the TIFF 6.0 extension is used. This is the tag 'DocumentName'. The TIFF information fields used by the American Memory project basically document the formal characteristics

[126] The website of this collection can be found at: <http://memory.loc.gov/ammem/aamhtml/aamhome.html> [cited 24 February 2004].

[127] The extraction of the TIFF tags is achieved with the 'TIFFdump' command of the LibTiff toolkit. See: <http://www.libtiff.org> [cited 23 Febuary 2004].

[128] Relevant information on the digital master files is not included in the information fields of the TIFF files and can be found in the background information on the website. Information on this sample collection can be found at: <http://memory.loc.gov/ammem/aamhtml/aambuild.html> [cited 12 February 2004].

of the digital master file. Formal characteristics of the original photograph are not stored in the information fields, such as the name of the photographer who created the original. The bibliographic data on the original is stored in a separate system.

As the archival master images of the photographs that are part of the American Memory collection are digitised with a fixed raster size, the physical dimension of the original on which the digital image is based cannot be derived from the relation between the 'ImageWidth' and the 'XResolution' values in the information fields of the TIFF file.

4.1.3 Conclusion

The TIFF 6.0 image file standard 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. Support for storage of preservation metadata in baseline TIFF is, however, fairly limited. Only a limited number of information fields are available for the storage of preservation metadata. Furthermore, the scope and purpose of these information fields is rather vaguely described. The requirements of the information field 'ImageDescription', for instance, are not specified. The coding of all significant characteristics of colour photographs is also problematic, because baseline TIFF does not support accurate coding of colour information. The TIFF 6.0 extension contains a high-quality device-independent colour space (CIELAB) that supports the accurate coding of colours. The digital master images of the American Memory project meet the standard baseline TIFF 6.0 specification.

4.2 Experiment 2: Durable digital image bitstream

The fundamental unit of digital data is the binary digit or bit that can be in either one of two states. A bitstream is a string of binary digits that make up a digital object such as a digital raster image. Digital data is not normally stored as a single, uninterrupted string of bits for three major reasons: ease of conceptualisation by the person who is developing ways to process the data, ease of specification and coding of the algorithms that will be programmed to control the computer, and efficiency of the algorithms ([TAN98] p. 319). The internal organisation of the data represented in the bits contained in the data file – the data structure – must be understandable by a program. The specification of the data structure prescribes the organisation of the bitstream. The TIFF standard specification, for instance, prescribes the organisation of a TIFF-formatted image and dedicated software is required to process the digital object. For example, a digital image-viewing program that understands the data structure of the TIFF file format is required to render a digital image on a computer screen. The data structure of digital objects is software dependent and software in its turn is hardware dependent. The threat that software and hardware can become obsolete is a constraint on the durability of the digital object.

The goal of this experiment is to assess the possibility of preserving the bitstream that forms a digital image by using a non-proprietary data structure. It is assumed that a bitstream preserved in a durable data format will result in a durable object. It is also based on the assumption that the bits that form the bitstream will remain physically intact in the long term. This is achieved by monitoring the storage medium of the bitstream and periodically refreshing the storage medium. Bitstream preservation results in long-term access to the authentic bits that make up a digital object as well as the possibility to carry out its originally intended function. Essentially, the bits that make up a raster image represent the arrangement and colour of the individual pixels that make up a rendered image. Also, metadata can be part of the bitstream. A large number of raster file formats do exist, but the rendering of the pixels that are part of the raster is dependent on a specific computing platform. A durable raster file format is platform independent, even in the long term.

Provided the bitstream is physically intact and can be read from the media, the main problem to be solved is the correct interpretation of the bitstream. The extraction of information that is in the bitstream is the most challenging problem to be addressed. For this, a durable data structure of the bitstream is required as well as a means of interpreting the data in a rational way.

4.2.1 XML as container for a durable bitstream

To a large extent the dependence on specific software and hardware for the processing of digital objects can be avoided by using the 'eXtensible Markup Language' (XML).[129] This standardised data format is considered as self-descriptive and does not require proprietary software to get access to the data. XML-formatted documents consist of mark-up codes and data. The mark-up codes describe the storage layout and logical structure of a document. XML provides a mechanism to impose constraints on the storage layout and logical structure of the data.

A simple text viewer is sufficient to take cognisance of the XML mark-up codes and the data between the mark-up codes. XML-formatted data is both human and machine interpretable. An essential requirement for gaining access to an XML-formatted bitstream is that the processor must understand the internal character encoding of the XML file. The autodetection or bootstrap mechanism of the XML encoding declaration is based on the ISO/IEC 10646 standard ([ISO10646:1993]).[130] This standard contains the legal binary codes that can be used for the mark-up codes or character data in the XML file. The XML mark-up codes are restricted in position and content to enable autodetecting the applied character encoding

[129] See: <http://www.w3c.org/xml> [cited 12 March 2004]. For a discussion of the durability issues of the XML data format, see section 2.2.5.

[130] XML processors must be able to read data formatted in the UTF-8 and UTF-16 encoding standard. The XML language requires that the encoding scheme used be declared in an XML file. UTF-8 and UTF-16 are specified in the annex of the ISO/IEC 10646 standard. Other encoding standards can be declared in an XML file, but they are not supported by default.

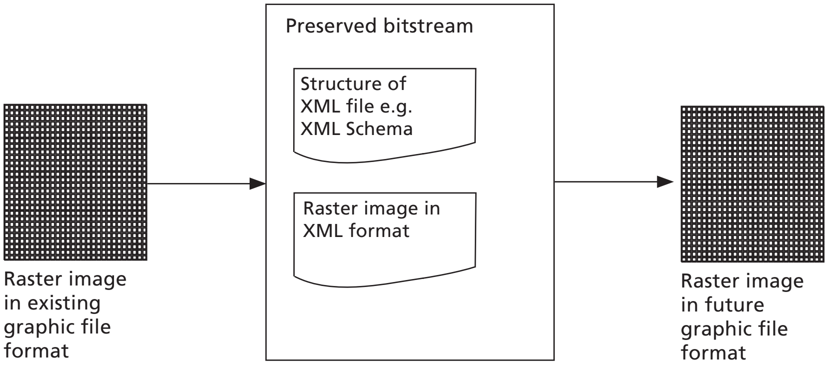

Preserved bitstream
Structure of XML file e.g. XML Schema
Raster image in XML format
Raster image in existing graphic file format
Raster image in future graphic file format

Figure 4.2 Components of a preserved bitstream in XML format

that is in use in common cases. According to the standard, the first sequence of characters of an XML file must be `<?xml`. This sequence of characters must be detected by the XML processor. After reading two or four bytes the processor will be able to understand the byte ordering method and thus solve the bootstrapping problem.[131]

A durable bitstream representing a digital raster image in XML data format consists of at least two parts:

– A valid XML-formatted file that contains the position and colour of the pixels that make up a raster image as well as the metadata that is required for the correct interpretation of the raster file. Valid implies that the XML file conforms to its content model. This is an instantiation of an object according to a predefined structure. This structure is also expressed in XML and this is the second part of the preserved bitstream in XML format.
– An XML-formatted file that contains the content model or definition of the structure of a valid XML-formatted raster image. For this, a definition language is required such as XML Schema. A definition language defines the specific vocabulary and specific hierarchical structure of an instantiation of an XML document.

Figure 4.2 illustrates the components of a preserved bitstream representing a digital raster image in XML format. A raster image in an existing graphic file format is translated into XML. The structure of this XML file is defined in a separate file, for instance an XML file, in conformance with the XML Schema language.[132] The combination of both files constitutes a durable bitstream. In order to render the pixels that make up the digital image on a computer screen or on paper, the raster image in XML format has to be converted into a binary graphic file format. A binary graphic file format is the data structure of a digital raster image that can only be interpreted by a dedicated program.

[131] See also section 'Autodetection of character encodings' in [XML04].
[132] See: <http://www.w3c.org/XML/Schema> [cited 34 March 2004].

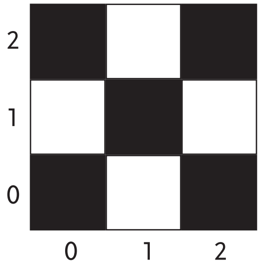


Figure 4.3 Bi-tonal bitmap consisting of nine pixels

Pixels formatted in XML

This section contains an illustration of a durable bitstream representing a digital raster image in XML format and introduces the requirements for the conversion of the existing digital raster image in binary format into XML format and the conversion of the XML-formatted bitstream into a (future) graphic file format. A digital raster image visible on a computer screen or printed on paper can be considered as a flat, table-like raster with a linear structure. The pixels are represented by the intersections of the horizontal and vertical lines of the raster. Each pixel has a colour code according to a specific colour model that is supported by the image file format, for instance RGB or CIELAB.

Figure 4.3 contains a small part of a simple bi-tonal raster image with only two colours: black and white. The numbers in the figure represent the horizontal and vertical position of the pixels in the bitmap. Depending on the user's view of the raster file structure, the raster file of Figure 4.3 can be expressed in the XML data format in a number of ways. One possible view is given in Figure 4.4. This figure contains a small part of an XML file that describes the simple bitmap depicted in Figure 4.3. The definition of the structure and content of the XML file can also be

```
<bitmap>
    <pixel>
            <position>
                <horizontal>0</horizontal>
                <vertical>0</vertical>
            </position>
            <colour>black</colour>
    </pixel>
    <pixel>
            <position>
                <horizontal>0</horizontal>
                <vertical>1</vertical>
            </position>
            <colour>white</colour>
    </pixel>
    ...
    ...
</bitmap>
```

Figure 4.4 Bitmap of Figure 4.3 expressed in XML

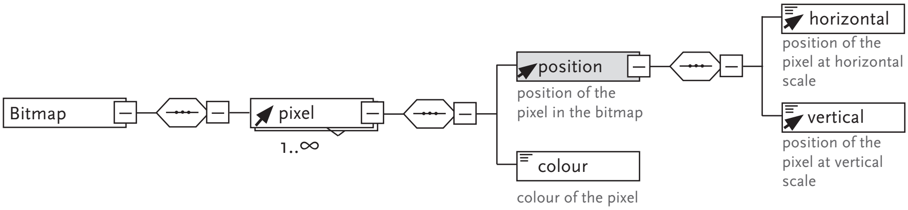

Bitmap
pixel
1..∞
position
position of the pixel in the bitmap
colour
colour of the pixel
horizontal
position of the pixel at horizontal scale
vertical
position of the pixel at vertical scale

Figure 4.5 Graphical representation of XML Schema representing the structure of the XML file of Figure 4.4

expressed in XML, for instance with the help of the XML Schema language. Figure 4.5 contains a graphical representation of the XML schema that contains the definition and structure of the XML file of Figure 4.4.

Requirements of durable bitstreams in XML format

Section 4.2.1 illustrates that the XML data format can be used to create a machine and human readable coding format for digital raster images. The XML data format is a 'self-documenting' data format, enabling the expression of the structure of a digital image. The following three functions are important regarding the creation and usage of a bitstream representing a digital image expressed in the XML data format:

– *Expression of the content model in XML*. The determination of the elements and attributes that must be part of the XML content model is important. Figure 4.4 and Figure 4.5 contain a simple, arbitrary example of an image file expressed in XML. The quality of the image file formatted in XML improves if the content model on which the XML instantiation is based meets the features of a standardised image file format.[133]
– *Binary to XML conversion*. This is the conversion of the binary image format into XML format. The binary image format is the specific data structure for which a dedicated program is required in order to render the data object. A generator is required that converts the binary image, for instance in TIFF format, into an XML-formatted file. This conversion process must of course take a content model into consideration. The content model must cover all features of the binary image file format.
– *XML to binary conversion*. This is the conversion of the XML-formatted bitstream representing the digital image into a binary format that can be processed by a computer platform, even in the future. The archived, original binary format needs to be re-generated since it is the only understandable

[133] The requirements of the image file format depend on the features of the original source and the intended use of the image. The determination of the file format is part of the benchmarking process. In a lot of cases the TIFF image format (see section 4.2) is a good candidate.

and usable format for the final user. The transformation of the pixels formatted in XML into a binary image can be considered as a problem of the future, because we have no information on the future data structures and hardware and software platforms. As it is assumed that the XML file can be opened and processed on all future platforms, at a future time a dedicated conversion procedure will need to be developed in order to render the pixels into an image. Currently available generic solutions to this problem are assessed in the following subsections.

At present there are three methods available that use the XML data format to create application-independent multimedia objects. The three methods are described, evaluated and compared with respect to their value for creating durable bitstreams. The evaluation will reveal to what extent building blocks are currently available that facilitate the creation of durable digital images in XML format. The three methods are:

– *The BSDL (Bitstream Syntax Description Language) framework* [AMI02]. The aim of the framework is to describe the structure of a bitstream in XML in order to facilitate the dynamic transformation of a digital object for multiple output specifications.
– *The UVC (Universal Virtual Computer) machine language* [LOR01], [LOR02]. The processing specification to render a data file, for example a digital image, is expressed in a machine language: the UVC machine language.
– *The Flavor (Formal Language for Audio-Visual Object Representation) language* [ELE03]. Flavor was developed to describe the bitstream syntax of multimedia objects. XFlavor is an extension that provides XML features by transforming the bitstream representation into a corresponding XML representation.

The three methods are described according to their ability to enable the creation of a durable bitstream of a specific digital object – a digital surrogate of an historical photograph.

4.2.2 Bitstream Syntax Description Language (BSDL)

BSDL was developed to enhance interoperability between different multimedia formats. Long-term access to digital objects or durability of digital objects is not mentioned as a rationale for the development of BSDL. The purpose of the schema language BSDL is to describe the structure of a bitstream. One of the main benefits of BSDL is that objects expressed in XML according to BSDL are easily accessible by text retrieval engines, and multi-layered multimedia objects, such as multi-resolution JPEG2000 images, can be manipulated on an individual layer basis. The XML representation of the bitstream enables the definition of editing operations on the XML document, for instance by modifying some element or attribute values.

BSDL is a generic approach by providing a method based on the XML data format for manipulating bitstreams. For each binary format, a specific content model is required and adapted for the specific features of the binary format.[134] To a certain degree the content model of Figure 4.4 can be considered as a BSDL representation of a proprietary simple image format. A difference compared with the examples in [AMI02] is that there is neither a binary source file nor a formalised file format on which the content model is based. Thus, the application of BSDL requires thorough knowledge of the multimedia format with respect to the way the bits are organised.

The processing instructions for creating a binary object from the XML representation are stored in a file using XSL (eXtensible Stylesheet Language).[135] XSL can be used for the definition of transformation rules and is expressed in the XML syntax. BSDL uses the W3C language XSLT to transform the XML descriptions of multimedia formats. This transformation does not result in a binary object but in a 'transformed XML description' and must be converted into a binary bitstream (see Figure 4.6).[136] An XSLT style sheet contains one or several templates defining the modifications to be applied to the elements or to the attributes matching a set of conditions. In this way, for instance, a digital image file format with multi-resolution images, such as the JPEG2000 data format, can be manipulated to retrieve a specific resolution. The result of the XSLT transformation is also an XML file 'from which it is possible to re-generate an adapted and compliant bitstream from the resulting description' (ref. [AMI02]).

[AMI02] describes the architecture of a system that converts a multimedia object into the XML data format according to a relevant style sheet expressed in XSLT. The system is also able to return a binary bitstream according to the specifications of a multimedia object.[137] A bitstream of a digital image consists of a structured sequence of binary symbols, specific to the coding format. The high-level structure of a bitstream is coded in the XML data format. The resulting XML document is considered as a 'bitstream description'. It does not describe the bitstream on a bit-per-bit basis, but addresses its high-level structure, for instance how the bitstream is organised in layers and packets of data. It does not deal either with the semantics of the bitstream, but only considers it as a sequence of binary symbols.

The value of the BSDL approach for creating durable digital surrogates of historical photographs can be determined when a digital image file format used for digital master images (for instance, the TIFF 6.0 image file format) is expressed

[134] [AMI02] contains a BSDL Schema for the following multimedia formats: a JPEG2000 image, an MPEG-4 video stream and a proprietary video stream.

[135] The World Wide Web consortium manages the XSL 'family' of style sheet languages (see: <http://www.w3c.org/Style/XSL>. Some style sheet languages are standardised such as XSLT (XSL Transformations).

[136] The image in Figure 4.6 is taken from [AMI02].

[137] The system architecture described in [AMI02] is based on the Cocoon Servlet provided by the Apache project, see: <http://xml.apache.org/cocoon2> [cited 4 June 2004].

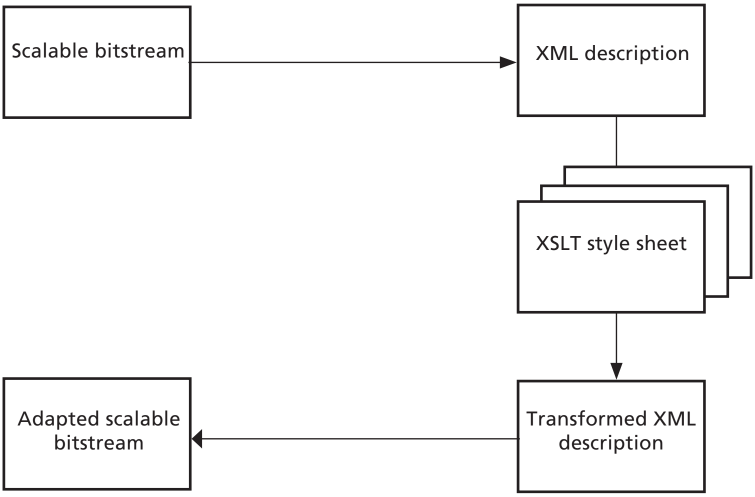


Figure 4.6 Transformation of a bitstream to the XML data format: the role of XSL

in the BSDL structure and when the system, as described in the system architecture, is implemented. Currently, this is not the case. An important aspect for the determination of the longevity of the BSDL approach will be the robustness of the system architecture and the technology used to implement the system. An assessment of the appropriate working of the system on future hardware and software platforms must be carried out in order to determine the durability of the implementation of BSDL. A strong point of BSDL is that it provides a framework for the expression of a binary data format resulting in a durable expression of this format. The absence of 'XML to binary' and 'binary to XML' functionality is a drawback of this method.

4.2.3 Universal Virtual Computer (UVC)

The goal of the UVC method is to provide a solution to the problem that digital data files and computer programs can no longer be used in the future. As a digital image is a data file and not a computer program, only the preservation of bitstreams representing data files with the help of the UVC method is taken into consideration.The basic idea of the UVC method is that the bitstream representing the data object is stored together with a logical view of the data. A logical view of the data is a view of the data that is easy to understand because it follows the way the user normally thinks about the data, rather than the internal representation often designed for efficiency ([LOR02] p. 35). Figure 4.4 contains a logical view of a raster image based on the principle that a raster image consists of pixels that have a place and a colour. In [LOR01] another logical view of a raster image can be found. In this logical view the coordinates of the pixels in the bitmap are not explicitly given.[138]

Not only the logical view but also the specification for processing the data on a future platform is archived. The process specification and the logical view definition are archived with the data. The processing specification is based on a Universal Virtual Computer (UVC). The UVC program is independent of the architecture of the computer on which it runs. A UVC interpreter has to be written for any future target machine. This means that every machine manufacturer will have to produce a UVC interpreter. The instruction set of the UVC is kept to a minimum. For the PDF data format, a proof of concept of the UVC method is published in [LOR02].Lorie acknowledges that data structures can be very complex ([LOR01] p. 347). The UVC examples given in the publications can be considered as simple data structures. It is not clear whether for complex data structures logical views can be designed that contain all relevant features. The logical view of a data format is stored in a schema that contains the structure of the data format. As there are many data formats, a lot of schemas will exist. Lorie introduces a generic 'schema to read schemas' approach. This schema to read schemas should be simple and intuitive and should endure for a long time to come, and be published in many places so that it remains known ([LOR00] p. 348). Lorie indicates that decoding rules for the data are simple and will therefore easily survive for a very long time ([LOR00] p. 350). The schema of a collection of images given as an example in [LOR01] is not detailed or accurate enough to meet high-quality criteria for digital master images. The colour codes, for instance, do not refer to objective colour code standards.

Lorie uses the proprietary DTD (Document Type Definition) construct for expressing schemas that contain the structure of the data object. In terms of durability, the XML Schema construct would be better as it uses the XML data syntax. The fact that only a short time ago DTDs were the norm for expressing the structure of a digital document is a warning. There is also a risk that XML Schemas will be superseded by some other method. To a certain extent, this is already the case because RDF (Resource Description Framework) is gaining ground. RDF is strong in expressing semantics but XML Schema supports explicit structural, cardinality and data typing constraints. This makes XML Schema a good candidate to use as a method to create durable bitstreams. The responsibility for extracting the logical data elements from the data stream lies in the methods. Lorie states that the only standards needed today are the UVC and the data model for data and metadata. They are simple enough to endure.

[138] In [LOR01] the logical view of a digital image consists of the colour values of dots. The number of lines in the image is part of the logical view, as well as the number of dots per line and the line number. It is not clear whether the rendering of the dots should start at the top or the bottom of the screen or paper.

The most challenging issue in using the UVC method to create a durable bitstream of a digital surrogate of an historical photograph is the creation of a logical view in XML format that contains all relevant details. Perhaps BSDL can be used to create high-quality XML representations of digital images. The UVC interpreter converts the XML-formatted data into a binary object. The UVC is used to archive methods that interpret the stored data stream ([LOR01] p. 352).

4.2.4 Formal language for Audio-Visual Object Representation (Flavor/XFlavor)

XFlavor [HON02] is an extension of the formal language Flavor [ELE03], which stands for 'Formal Language for Audio-Visual Object Representation'. Flavor was developed for the description of a bitstream syntax representing multimedia objects.[139] It was developed as the means to represent multimedia data in order to simplify and speed up the development of software that processes multimedia data. The purpose of converting multimedia data that is available in binary format into an equivalent XML document is for easier and more flexible manipulation of the data. With the document, the bitstream layer is abstracted from applications and the semantic values of the data (for instance, width and height of an image) are directly available, whereas with the bitstream format such values must be extracted via bit string manipulations, according to the Flavor syntax. Another advantage of using XML representation of data is that software tools are provided with generic access to multimedia data (usually with a generic XML parser). Durability of digital objects is not mentioned as a design objective for Flavor and XFlavor.

Flavor can be used to describe and process bitstreams according to a formalised, specified syntax. Flavor can also be used as a media representation documentation tool and as a method for redefining the syntax of content in both forward- and backward-compatible ways. The separation of parsing from the remaining coding/decoding operations allows its complete substitution as long as the interface (the semantics of the previously defined fields) remains the same.

In order to simplify interoperability among different applications, XFlavor was developed. XFlavor has three major functions: (1) Given a Flavor description of the bitstream syntax, a corresponding XML Schema can be created; (2) given a Flavor description of a bitstream syntax, XML document generating code can be created; (3) conversely, for an XML representation, an equivalent bitstream can be generated. XFlavor offers an alternative form of multimedia data so that different and interoperable applications can be created for the same data even if the expected syntax of the data is different among the various applications.

General-purpose languages such as Java and C++ do not provide native facilities for coping with bitstream-oriented data. A possible application of XFlavor is content extraction. XML-formatted elements are easy to retrieve by indexing

[139] Detailed information on Flavor as well as the source code, the Flavor translator and the run-time library can be found at: <http://flavor.sourgeforge.net> [cited 29 March 2004].

tools. XML-based applications are able to support different bitstream syntaxes, as long as the required elements are available in the document and the tags remain the same. XSLT can be used for transforming XML documents. With XSLT, any application can be enabled to use any XML document as long as the actual data in the document is useful in the transformation. This 'universal interoperability' is achieved by transforming XML documents into structures tailored to different applications. A transcoder can easily be created by simply declaring a set of rules for transforming the source elements. The style sheet (containing XSLT rules), along with the source document, can be fed into an XSLT processor to generate a new document with the desired structure and information. XFlavor has the ability to transform a given XML representation of multimedia content back into the bitstream representation. Converting the XML representation of the multimedia data back into the binary representation is straightforward, as all the information for constructing the bitstream is provided by the attributes of each element.

The conversion of the raster image file into the XML data format as presented in Figure 4.4 is based on an ad-hoc interpretation of the raster file. The content model is very simple, as the structure of a graphic file is much more complicated. The XML Schema (and thus the XML instantiations that are derived from an XML Schema) must be much richer and more detailed. With Xflavor, it is possible to convert all details of a graphic file format into the XML data format. The XFlavor application contains a procedure for converting images formatted in the GIF image format into the XML data format. XFlavor creates an XML Schema that contains the structure of the GIF image file format.

In order to minimise the cost of storage and transmission, digital audio-visual objects, such as digital images, are coded in a specific way. Encoders and decoders, as well as applications, must interpret the specific way in which the bitstream is organised in order to be able to render the object. The data to be represented is converted to a sequence of binary values of arbitrary lengths, according to a specified syntax. In order to simplify and speed up the development of software that processes coded audio-visual data, generic media representation languages have been developed. These languages are based on the XML data format. Due to the usage of the XML data format, digital preservation can benefit from the development of these generic media representation languages. The XML data format is platform neutral and the XML-formatted objects can outlive the original binary digital object. Once a digital object is available in the XML data format, with the help of XSLT it can be formatted in a large number of other data formats.

Flavor allows the formal description of the bitstream syntax. The description is based on a well-defined grammar. Several existing bitstream formats have already been described in Flavor. Flavor provides a formal way of specifying how data is laid out in a serialised bitstream. It is based on the principle of separation between bitstream parsing operations and encoding, decoding and other operations. Flavor is a media representation language.

Table 4.3 Comparison of three methods with respect to 'bitstream preservation'

	Binary to XML conversion	Content model in XML	XML to Binary conversion
BSDL	not implemented	available	not implemented
UVC	Available	partly available	available
XFlavor	partly available	available	partly available

XFlavor is a translator that generates XML Schemas from Flavor descriptions. Additionally, a method is available for producing an XML document from a Flavor described bitstream. As a complement, a software tool for converting the XML document back into the original bitstream format is also provided as part of this Flavor package.

The Flavor/XFlavor method is the most complete method available for creating usable and durable bitstreams. Usable bitstreams: because facilities are available to convert a binary data format into the XML data format and to convert an object in XML back into a (future) binary format. Durable bitstreams: because the non-proprietary XML data format is used to code the bitstream. The Flavor language creates Java and C++ code that can process the bitstream. The durability of the C++ and Java platform is the central issue in comparing the Flavor method with the UVC approach. The UVC approach is based on the principle that a virtual machine instruction set is the core of the method. Multi-platform, multi-user languages such as Java and C++ are not considered durable by the UVC approach.

4.2.5 Evaluation of bitstream preservation methods

Table 4.3 gives an overview of three bitstream preservation methods – BSDL, UVC and XFlavor – regarding their ability to perform three essential tasks in creating a durable bitstream representing a digital raster image. These tasks are: (1) the conversion of the proprietary binary data format into the XML data format; (2) the expression of the content model of the XML file, and (3) the conversion of the XML data format into a (future) proprietary binary data format. The table gives an indication of the quality of the task as part of a method. The table indicates whether a task is available in the method. A task can also be partly available. This means that the method can perform the task with digital objects in general but that adjustments to the method are required to enable the processing of digital raster images.

The publications available on the BSDL method only contain a design of a system architecture that is able to carry out the conversions. All essential features of a data format are part of the content model that is expressed as an XML Schema. The UVC method has the best approach regarding the conversion between XML and proprietary binary data representations of digital objects, because the virtual computer will be able to process the data, provided that a UVC emulator is available on all future computer platforms. As the UVC is intentionally kept very simple, it is assumed that it will not be difficult to create this emulator. For each data format a content model has to be created.

Flavor/XFlavor has already implemented all the necessary functionalities, but is based on the Java and C++ language. The stability and durability of these languages is the most important factor in the assessment of the method for the conversion between a binary and an XML data format.

Importance of based on a formal standard

Both the UVC and the XFlavor methods will be appropriate bitstream preservation methods for raster images. The durable formulation of the logical view (a view of the data that follows the way the user normally thinks about the data) of a digital object is not as easy as it seems at first glance. Even with tangible objects, such as books, it can be difficult to create a generally accepted logical view. For digital objects this is even more difficult. In any case the ability of a method to represent the logical view of the digital object will be an important factor for the degree of success of a given method. The logical view makes the object understandable in an obvious way for the user.

The logical view implies the interpretation of the representation of the data file, in this case a digital raster image. As humans do the interpretation, the chances are that in the future people may not understand the interpretation that was created in the past, or that essential features of the digital image are not included in the logical view. This can be illustrated by the following example. A high-quality digital master image serving as a surrogate of a colour photo requires the application of a device-independent colour space, such as the CIELAB standard. Standard RGB colour coding is sufficient for the creation of reference images on a computer screen. But a faithful reproduction on paper, for instance, requires the device-independent coding of the pixel colours. The logical view of the bitstream expressed in the XML data format of a high-quality digital surrogate should be based on the features of a standard file format that meets all the requirements of a digital master file, including issues such as device-independent colour coding schemes.

4.3 Experiment 3. Durable formulation of preservation metadata

The aim of the third experiment is to examine procedures for the explicit, unambiguous and standardised expression of metadata schemas that serve as preservation metadata. Good-quality metadata is an important facilitator for the durability of the digital objects that are described with it. Depending on the designated community a number of metadata schemas can be distinguished. This is illustrated by Table 2.2.

Designated communities tend to ‘mix and match’ metadata elements, resulting in application profiles, as is illustrated in Figure 2.5. This makes the durable formulation of preservation metadata not only a matter of compiling the most appropriate set of metadata elements but also a matter of expressing these metadata elements in a clear, standard way. If the terminological principles and methods for expressing metadata are standardised, it is possible to understand and assess the value of the digital objects that are documented with it, now and in the future.

One of the observations made in section 2.2 of this study is that the unambiguous understanding of the semantics of metadata schemas is highly problematic. This is mainly due to the fact that metadata elements expressed in natural language are often formulated intuitively and can be interpreted in a number of ways. This can be illustrated with an example that is relevant for the subject of this study. The metadata element with the label 'date', as related to a digital surrogate of an analogue photograph, can have a number of connotations. Is it the date the digital image was created? Is it the date the depicted scene occurred? Or is it the date the original photograph was taken? Also, the format of the metadata element can be expressed in a number of ways.[140]

Ultimately, a metadata registry, an information system for registering metadata, is needed to optimise the formulation and reuse of metadata elements, but permanent metadata registries for memory institutes do not exist at the moment. Some promising initiatives do exist but a main drawback of these registries is that they lack a common way of expressing the features of metadata elements. This is illustrated by Table 2.3, which contains information on the structure of three metadata registries. Thus, good practice in declaring metadata schemas and metadata elements is an important starting point towards the establishment of operational metadata registries in the future. The quest for a standard to formulate metadata elements in an unambiguous way is described in section 3.2 of this study and leads to the observation that the ISO/IEC 11179 family of standards regarding the formulation of Data Elements and Metadata Registries is relevant for the unambiguous formulation of metadata schemas.

Focus on Technical Metadata

The experiment elaborated on in this section investigates whether a specific metadata schema relating to the technical aspects of digital surrogates of historical photographs can be formulated according to the ISO/IEC 11179 family of standards. Section 2.2.3 of this study makes it clear that technical metadata for digital objects is important for digital durability. The set of metadata elements that are part of the *Data dictionary – Technical metadata for digital still images* ([NISOZ39.87:2002]) is an emerging standard for the formulation of technical metadata.[141]

The members of the committee that compiled the list of metadata elements of the [NISOZ39.87:2002] standard represent prominent Anglo-American memory institutes. The committee is active within the framework of NISO and AIIM, two important standard creation bodies. This makes [NISOZ39.87:2002] relevant

[140] The DCMI Date Working Group of the Dublin Core Metadata Initiative is dedicated to discussing issues related to the representation of date and time in metadata. See: <http://www.dublincore.org/groups/ date/> [cited 14 December 2004].

[141] For a detailed description of [NISOZ39.87:2002], see page 62 *et seq*.

for the compilation of preservation metadata that improves the durability of digital images. The description of the metadata elements in the [NISOZ39.87:2002] standard document is done in an ad-hoc manner. The terminology used is defined in the [NISOZ39.87:2002] standard document. Also, the attributes of the metadata elements are clarified. They are named 'metadata fields' in the [NISOZ39.87:2002] standard document.

This section consists of two parts. In the first part relevant information concerning the process for formulating metadata elements according to the ISO/IEC 11179 family of standards is presented, whereas in the second part of this section an existing set of metadata elements is examined. This examination is based on the procedures described in the first part of this section and can be considered as an experiment, because the theoretical model-type approach is applied in practice.

4.3.1 Procedures for the durable formulation of Data Elements

According to the ISO/IEC 11179 family of standards, a Data Element is a unit of data for which the definition, identification, representation and permissible values are specified by means of a set of attributes. A Data Element must be documented with an administration record. The administration records of Data Elements are important for digital durability, because they clearly define the attributes of digital objects and this improves the quality of the metadata. For the purpose of this experiment, the Data Element as defined by ISO/IEC 11179 is considered equivalent to metadata elements that are part of a metadata schema.

The ISO/IEC 11179 family of standards deals with the managing of the semantics of data and specifies the structure of a metadata registry in the form of a conceptual model, not intended to be a logical or physical data model for a computer system. The procedures for the consistent registration of Data Elements and their attributes in a registry are described in the Technical Report: *Procedures for achieving metadata registry content consistency – Part 1: Data elements* [ISO20943-1:2003]. The procedures for the consistent registration of Value Domains are described in the Technical Report: *Procedures for achieving metadata registry content consistency – Part 3 – Value domains* [ISO20943-3:2004].[142] A Value Domain is a set of valid values for one or more Data Elements. The relation between Data Elements and Value Domains is illustrated in Figure 3.2.

In the design phase of metadata elements, decisions have to be made regarding the use of the Data Element construct to formalise semantics or whether to use a Value Domain for this. This issue resembles the use of either elements or attributes when creating an XML Schema (see: [ANT04] pp. 27-29). The registration approaches described in the Technical Reports are biased towards the subsequent subjects of the reports.[143] For the purposes of the experiment, the choice is to give

[142] The Technical Reports contain 176 pages in total. In this section an outline of the content is given as far as it is relevant for the experiment; that is to say, the standardised registration of preservation metadata.

[143] The Technical Report on the registration of Data Elements states '... the choice will always

the Data Element concept higher priority than the Value Domain for the durable formulation of metadata.

The purpose of [ISO20943-1:2003] is to describe a set of procedures for the consistent registration of Data Elements and their attributes in a registry. It is not a data entry manual, but a user's guide for conceptualising a Data Element and its associated metadata items for the purpose of establishing good-quality Data Elements. An organisation may adapt and/or add to these procedures as necessary. Data Elements are ideally the result of a process of development, involving several types of abstraction, a well-developed tool for analysis and conceptualisation, common in data modelling techniques. The three types of abstraction of most interest to Data Element development are:

- *Specialisation / generalisation*: relationship between two classes, where all items in one (subclass) are also in the other (superclass).
- *Concatenation (or composition):* involves the development of composite values by concatenation of character sequences from source value.
- *Aggregation:* is used to express a relationship among Data Elements in which the higher layer describes a characteristic of a whole and the lower layers are factors affecting that characteristic.

Registration of a Data Element in a registry requires that certain characteristics of the Data Element be recorded to clearly describe and define it. These characteristics are stored as attributes of the Data Element. In the Technical Report [ISO20943-1:2003] two approaches to Data Element registration are distinguished:

- *The bottom-up registration procedure*: provides for the basic metadata attributes (for instance, definition, name and permissible values) about the Data Element to be completed prior to defining the conceptual information about the Data Element. A bottom-up approach for Data Element registration might be used where the registry is intended to serve as a distribution mechanism for metadata that describes the data in objects such as public data sets.
- *The top-down registration procedure*: the registration begins with the identification of Data Element Concepts.

Figure 3.2 illustrates the two approaches. The bottom-up registration procedure starts at the representational level and the top-down registration procedure is initiated at the conceptual level. The bottom-up and top-down procedures differ only in the order in which the practitioner analyses the Data Elements and formulates their associated items and attributes. Regardless of the approach, the same rules and guidelines apply to the associated metadata items and attributes ([ISO20943-1:2003] p. 6). Both approaches are described below.

be to treat the [code] sets as Data Elements unless explicitly stated' ([ISO20943-1:2003] p. 1). The Technical Report on the registration of Value Domains states '... the choice will always be to treat the [code] sets as Value Domains, unless explicitly stated' (ISO20943-3:2004] p. 1).

Table 4.4 Eight steps in the bottom-up registration process of Data Elements

Step	Action
1	Formulate the definition of a Data Element
2	Identify the permissible values and Value Domain
3	Enter the representation class
4	Assign the Data Element name and Data Element identifier
5	Record other Data Element attributes
6	Specify the Data Element Concept and Conceptual Domain
7	Record classification scheme attributes
8	Assign the registration and administrative status

Bottom-up approach to Data Element registration

An appropriate bottom-up approach to Data Element formulation and registration may be to work from Data Elements to Data Element Concepts (see Figure 3.2). The Data Element 'date', for instance, may be related to the Data Element Concept 'the date the digital surrogate is created'.

According to [ISO20943-1:2003], the bottom-up registration process of Data Elements consists of eight steps that are described below. Prior to the bottom-up formulation of the Data Elements, a good understanding of the semantic content of the Data Elements is essential. Data Elements can be related to an enumerated or non-enumerated domain, or be part of an international standard or an information system. Content research must make clear whether the Data Element is described in an existing international, national or organisational standard and whether the Data Element exists in a registry that has the potential for being reused.

The eight steps of the bottom-up procedure for registering a Data Element can be found in Table 4.4. The first step involves the capture of the essential semantic content of a Data Element in the form of a Data Element definition. Rules and guidelines for formulating the essential semantic content of a Data Element in a Data Element definition are given in [ISO11179-4:2004].

In the next step the permissible values and Value Domain for a Data Element are identified. Different attributes are used depending upon whether the permissible values are enumerated or non-enumerated. [ISO11179-3:2003] identifies the attributes that describe the domain of permissible values.

In the third step of the bottom-up registration process, the representation class of the Data Element is entered. The representation class describes how the Data Element is represented. A set of classes makes it easy to distinguish between the elements in a registry. Examples of representation classes are 'amount', 'number', 'date' and 'time'. A representation class is an optional attribute of a Data Element.

Next, a name and identifier are assigned to a Data Element. Part 5 of the ISO/IEC 11179 family of standards gives principles for naming and identification of

Table 4.5 Eight steps in the top-down registration process of Data Elements

Step	Action
1	Specify the Data Element Concept and Conceptual Domain
2	Formulate the definition of a Data Element
3	Enter the representation class
4	Assign the Data Element name and Data Element identifier
5	Identify the permissible values and Value Domain
6	Record other Data Element attributes
7	Record classification scheme attributes
8	Assign the registration and administrative status

Data Elements.[144] The name of a Data Element very much resembles the Data Element definition as created in step 1 of the registration process. As the name of a Data Element is developed 'according to the naming convention for a particular name context' ([ISO20943-1:2003] p. 9), the requirements for the semantic value of the name are less strict than the Data Element definition of step 1. Each Data Element must have a unique identifier assigned by a registration authority that establishes its own identification scheme. Part 6 of the ISO/IEC 11179 family of standards elaborates on the formulation of identifiers. The requirements for identifiers as described in [ISO20943-1:2003] are less strict than the global identifier schemes discussed in section 3.3.1 of this study. Concerning the global uniqueness of Data Element identifiers, a unique Registration Authority Identifier (RAI) must be established giving the registration authority the freedom to use any identifier scheme for its Data Elements within the metadata registry.

In the fifth step of the bottom-up registration process, a number of optional Data Element attributes can be recorded in the form of a so-called profile. In addition to the definitional attributes and identifying attributes, there are administrative, relational, classifying and other miscellaneous attributes that serve to define and describe a Data Element. For the bottom-up registration procedure of Data Elements, four attributes can be recorded. First, attributes on the organisation that submits the Data Element to the registry; second, the stewardship contact that contains data on the organisation that has been delegated the responsibility for managing a set of data resources; third, explanatory comments that can be used to provide remarks about the Data Element that are not appropriate to include in the Data Element definition attribute; and fourth, data on the origin of information about the Data Element, for instance a standard, document, system or group.

In the sixth step of the bottom-up registration process, it is possible to specify conceptual information about the Data Element through the Data Element Concept. The Data Element Concept may relate several Data Elements that record

[144] This information is based on Technical Report [ISO20943-1:2003]. This Technical Report refers to part 5 of ISO/IEC 11179, which is 'to be published' ([ISO20943-1:2003] p. 2).

data about that concept with different representations. A Data Element Concept can be associated with many Data Elements, but is associated with only one Conceptual Domain (see Figure 3.2). Within the Conceptual Domain 'countries of the world', for instance, the Data Element 'Name of country' is a representation of the Data Element Concept 'Country identifier' and both the country identifier 'NL' and 'NLD' may refer to the Data Element representation 'The Netherlands'.

In the last but one step of the Data Element registration process, classification scheme attributes are recorded where applicable. Standard document [ISO11179-2:2000] describes general categories of classification, such as keywords, thesaurus terms and taxonomy data. Classification helps to add information not easily included in definitions, helps to organise the contents of a registry, and helps to provide access by supporting more meaningful queries.

The eighth and last step in the bottom-up registration process of Data Elements consists of the recording of the status of the registered Data Element.[145] Part 6 of the ISO/IEC 11179 family of standards specifies the layers of the registration status.

Top-down approach to Data Element registration

A top-down approach to Data Element registration is used 'where information available to the practitioner provides an overall understanding of the Data Element, including knowledge of its characteristics and relationships. The practitioner can then identify and define objects and properties upon which Data Element Concepts are based' ([ISO20943-1:2003] p. 44).

The top-down procedure for registering a Data Element consists of the same steps as the bottom-up procedure, but the steps are carried out in a different order (see: Table 4.5). In contrast to the general description of the bottom-up registration process of Data Elements, Technical Report [ISO20943-1:2003] covers the top-down registration process in the form of an example that is far from relevant for the subject of this study.[146] In the Technical Report the description of the top-down procedure is based on the example and contains a number of deviations from the top-down approach as it is introduced earlier in the Technical Report.[147] This makes it difficult to compile a general, abstract description of the top-down registration process.

[145] Examples of registration status are: 'incomplete' (when not all attributes for a Data Element are recorded), 'recorded' (when all mandatory attributes for a Data Element have been entered), 'certified' (when the Data Elements went through some quality review process), and 'standard' (when consistent representation and understanding of the data are achieved within the designated community).

[146] The example relates to professional organisations that wish to track the ability of experts to communicate in various languages.

[147] The description of the top-down approach to Data Element registration based on the example involves seven steps ([ISO20943-1:2003] pp. 44-45) and this is contradictory with the eight steps as introduced earlier in the Technical Report.

The top-down registration process might be based on an information system or a document that establishes the scenario for registering the Data Elements. The convention for establishing the context for names and definitions for Data Elements should take a number of issues into consideration, such as the scope of the Data Elements and the authority that is related to the Data Elements. The first step in the top-down registration procedure concerns the construction of object classes upon which Data Elements can be based. An object class is a type of administered item (see Table 3.3). If the object classes exist in the metadata registry, they should be used in the creation of Data Element Concepts. Where they do not exist in the registry, both name and definition should be entered. In the second step of the top-down approach to Data Element registration, the attributes of a Data Element are formulated using the rules and guidelines for Data Element definitions described in standard document [ISO11179-4:2004].

The next three phases of the top-down registration procedure involve the registration of the representation class, the formulation of the name of the Data Element and the specification of the Value Domain and permissible values. A Value Domain can be enumerated or not enumerated and this influences the registration of the permissible values. The next, sixth, step can be approached in the same way as the fifth step of the bottom-up registration process (see: Table 4.4). In the seventh step of the top-down registration approach the Data Element is classified and in the last step the registration status is recorded. These steps are in line with step 7 and step 8 of the bottom-up approach.

Registration of Value Domains

The procedures for the consistent registration of Value Domains in a registry are described in the Technical Report *Procedures for achieving metadata registry content consistency – Part 3 – Value domains* [ISO20943-3:2003]. A common understanding of the content of the Value Domain attributes enables the sharing of metadata between registries, despite their differences. A registry does not contain data itself. It contains the metadata that is necessary to clearly describe, inventory, analyse and classify data. It provides an understanding of the meaning, representation and identification of units of data.

A Value Domain is a set of permissible values. A permissible value is a combination of some value and the meaning of that value. The associated meaning is called the value meaning. A Value Domain is the set of valid values for one or more Data Elements. It is used for validation of data in information systems and in data exchange. Two types of Value Domains can be distinguished: enumerated and non-enumerated. An enumerated Value Domain is a Value Domain where all the permissible values are listed explicitly. A non-enumerated Value Domain is a Value Domain where the permissible values are expressed using a rule, called a non-enumerated Value Domain description. Thus, the permissible values are listed implicitly.

A Conceptual Domain is a set of value meanings. It is a concept for which the extension is a collection of Value Domains. Conceptual Domains, too, come in two main types: enumerated and non-enumerated. The value meanings for an enumerated Conceptual Domain are listed explicitly. A non-enumerated Conceptual Domain is expressed using a rule. The relation between Value Domains, Conceptual Domains, Data Elements and Data Element Concepts is illustrated in Figure 3.2 .

4.3.2 Standardised registration of metadata elements derived from 'Data dictionary – Technical metadata for digital still images'

In this section the ISO/IEC 11179 family of standards concerning the formulation of Data Elements is applied in a practical experiment by using the approaches described in the preceding section. More concretely, this section examines whether it is possible to formulate the metadata elements of the *Data dictionary – Technical metadata for digital still images* ([NISOZ39.87:2002]) in compliance with the ISO/IEC 11179 family of standards. The relevance of this experiment is that it makes clear to what extent metadata that is important for the durability of digital surrogates of historical photographs can be formulated in an unambiguous way.

It is likely that the method used by the NISO committee to compile the [NISOZ39.87:2002] data dictionary corresponds with the bottom-up approach (see Table 4.4). The 15 specialists of the committee are aware of the context of the problem area and are acquainted with relevant related standards and essential semantic content of the metadata elements; these requirements must be met prior to the bottom-up formulation of Data Elements. Although it seems plausible to apply the bottom-up approach, the eight steps of the bottom-up registration process of Data Elements (see Table 4.4) are not distinguishable in the [NISOZ39.87:2002] document and also the attributes of the Data Elements do not meet the requirements of the ISO/IEC 11179 family of standards. The administrative status of the Data Elements, for instance, is not explicitly formulated.

For the experiment, the existing [NISOZ39.87:2002] document is used as a point of departure and a top-down procedure is applied to register the Data Elements according to the ISO/IEC 11179 family of standards. After all, the top-down approach is applicable in a situation where an 'overall understanding of the Data Element including knowledge of its characteristics and relationships' is available. This overall understanding is available in the form of the [NISOZ39.87:2002] data dictionary document.

Scenario for top-down registration of preservation metadata

The following scenario illustrates the top-down approach to registering Data Elements relevant for memory institutes that, for digital preservation purposes, wish to track the technical metadata of digital still images. Based on the top-down approach for the registration of Data Elements as described in Technical Report

Table 4.6 Three approaches for top-down registration of Data Elements

	Approach 1	Approach 2	Approach 3
CD	All metadata schemas	Metadata schemas for digital preservation	Preservation metadata for digital still images
DEC	Metadata schemas for digital preservation	Metadata schemas for digital still images	Preservation metadata for digital still images according to NISO data dictionary for digital still images ([NISOZ39.87:2002])
DE	Metadata elements for digital still images	NISO data dictionary for digital still images ([NISOZ39.87:2002])	Categories of metadata elements of NISO data dictionary for digital still images ([NISOZ39.87:2002])
VD	Value Domains related to DE for digital still images (e.g. list of digital file formats, list of compression schemes, etc.)	Value Domains related to DE of NISO data dictionary for digital still images ([NISOZ39.87:2002])	Value Domains related to DE of categories of NISO data dictionary for digital still images ([NISOZ39.87:2002])

[ISO20943-1:2003], the specification of the Data Element Concepts and Conceptual Domains starts with a determination of the source and context of the Data Elements to be registered. The top-down approach begins by documenting the conceptual level of the metadata schema followed by the documentation of the representational level (see Figure 3.2). The context, or universe of discourse, determines the level of detail that is covered by the metadata.[148] Table 4.6 contains three possible approaches for the top-down formulation of Data Elements.

The Conceptual Domain (CD) of the first approach has metadata schemas in general as the semantic domain, whereas the related Data Element Concept (DEC) concerns metadata schemas specific to digital preservation. This DEC is expressed by Data Elements (DE) that address digital preservation of digital still images. Possible examples of Value Domains (VD) that represent Data Elements are lists of digital image file formats and lists of compression schemes for digital image files.

By restricting to metadata schemas specific to digital preservation, the CD of the second top-down approach to the registration of DEs has a more limited scope than that of the first approach. As a consequence, the DEC is also more specific, as are the DE and VD. The DEs relate to the metadata items that are part of the NISO standard. The CD of the third approach concerns preservation metadata for digital still images and has metadata items according to the NISO data dictionary for digital still images as its DEC. At the representation level, the DEs are categories of the NISO data dictionary for digital still images represented by VDs of these categories. The third approach is worked out further, because it is closely allied with the structure and content of the [NISOZ39.87:2002] standard document.

According to Technical Report [ISO20943-1:2002], the construction of a DEC is based on Figure 4.7: MIX XML Schema as a basis for a data model for preserva-

[148] ‘Context’ is a type of Administered Item as distinguished by the ISO/IEC 11179 family of standards, see Table 3.2.

tion metadata for digital still images relevant object classes. A data model[149] makes object classes and associated properties explicit.[150] The [NISOZ39.87:2002] document does not contain a data model. In line with the inductive research strategy of this dissertation, existing building blocks are used to answer the research question. The MIX XML Schema, which consists of elements based on the [NISOZ39.87: 2002] data dictionary, can be used for the construction of a data model.[151] An XML Schema defines the syntax of an XML-formatted document and the MIX XML Schema contains the rules for XML-formatted documents that contain technical metadata for digital still images. A part of the MIX XML Schema is given in Figure 4.7.[152] The 'MIX' element on the left should be interpreted as object class 'digital still image'. The metadata elements of the [NISOZ39.87:2002] data dictionary are classified in a number of sections and sub-sections. These sections and sub-sections can also be considered as object classes.[153] More information on the categories 'Basic image parameter', 'Image creation', 'Imaging performance assessment' and 'Change history' can be found on page 63. Not all items formulated in the [NISOZ39.87:2002] document are translated into XML Schema elements in a consistent way in the MIX XML Schema. In the MIX XML Schema, the element 'Host' is created containing a sequence of the data dictionary elements 'HostComputer', 'Operating System' and 'OS Version'. The element 'Host' can be seen in Figure 4.7. Also, in the MIX XML Schema the element 'Wrap' is created containing tags from [NISO39.87:2002] that give data on the colour map of the digital image. Tags relevant for Grayscale data in the still image are defined in the MIX XML Schema as sub-elements of 'GrayResponse'. Both the 'Wrap' and 'GrayResponse' elements are sub-elements of the element 'Energetics', which is visible in Figure 4.7. The [NISOZ39.87:2002] metadata element 'ImageIdentifierLocation' ([NISOZ39.87:2002] p. 12) in the MIX XML Schema became the name of an attribute and not the name of an element.

[149] A data model is a graphical and/or lexical representation of data, specifying its properties, structure and inter-relationships ([ISO11179-1:2004] p. 3).

[150] Annex D of Technical Report [ISO20943-1:2003]: 'Example of complete associated metadata item descriptions using a top-down approach to data element registration' ([ISO20943-1:2003] pp. 98-124) starts with a very simple data model with three object classes. The Technical Report uses 26 pages to document all items that are based on this simple data model. This is an indication of the extensive work that is required to formalise Data Elements in a metadata registry.

[151] MIX stands for: 'NISO Metadata for Images in XML Schema'. More information on the MIX XML Schema can be found on page 70. Figure 2.6 contains a part of the MIX XML Schema.

[152] The graphical representation of the MIX XML schema acting as data model in Figure 4.7 consists of a 'Schema Design View' of the XML development environment XML Spy, version 4.4.

[153] An object class is defined by the ISO/IEC 11179 family of standards as 'a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behaviour follow the same rules' ([ISO11179-1:2004] p. 10).

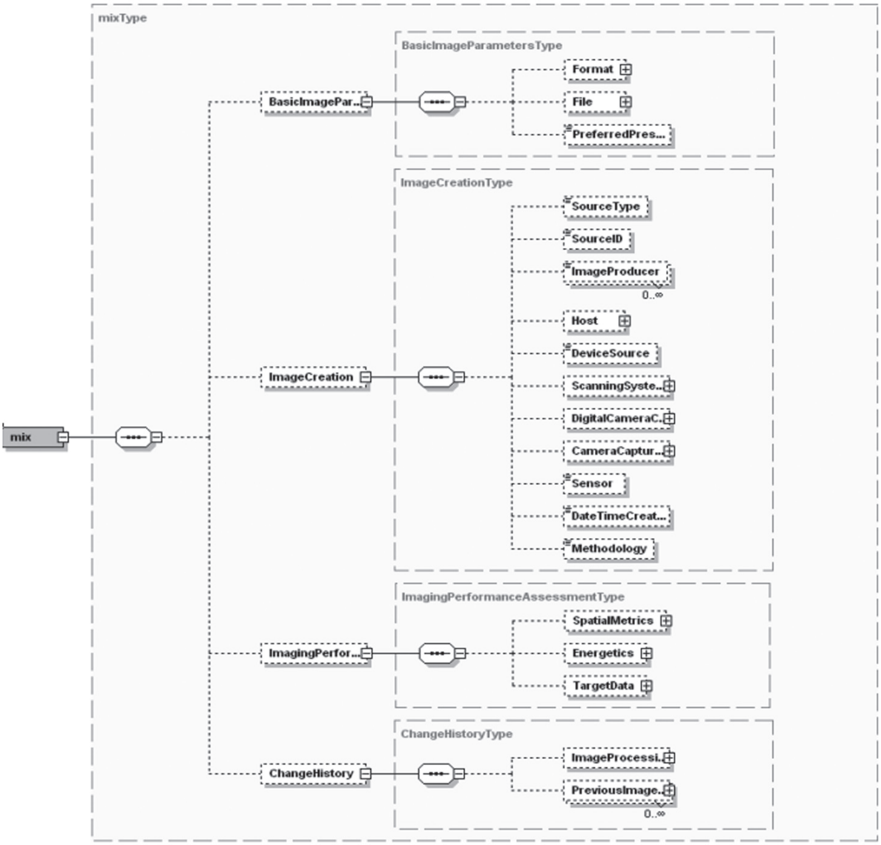


Figure 4.7 MIX XML Schema as a basis for a data model for preservation metadata for digital still images

The issues discussed above demonstrate that, due to the XML Schema syntax rules, the MIX XML Schema is more consistent and clearer than the [NISO Z39.87:2002] document. The comparison between the [NISOZ39.87:2002] data dictionary and the MIX XML Schema also proves that it is apparently inevitable that mutations and adjustments occur when a generic standard is made explicit in a data model. Each element of the MIX XML Schema contains an annotation tag. This annotation tag consists mainly of the 'definition' and 'notes' attributes of the [NISOZ39.87:2002] metadata elements with some minor adjustments. In the MIX XML Schema, enumerated elements are formulated in much more detail than in the [NISOZ39.87:2002] data dictionary. The data dictionary, for instance, sums up five MIME types associated with the image data, whereas the MIX XML Schema contains a list of 54 different MIME types.[154] The third approach to the top-down

[154] MIME stands for 'Multipurpose Internet Mail Extensions'; it extends the format of Internet mail to allow non-text messages.

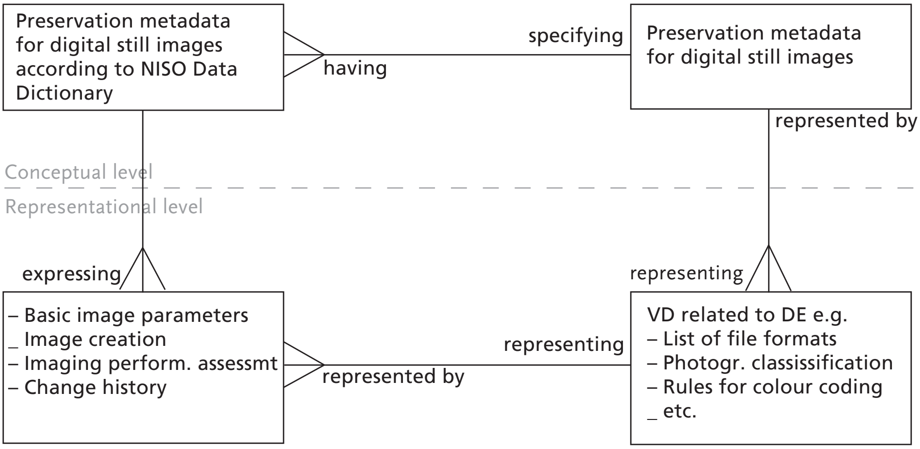


Figure 4.8 Metamodel of preservation metadata for digital still images based on [NISOZ39.87:2002]

registration of Data Elements as given in Table 4.6, combined with the data model of Figure 4.7, results in the ISO/IEC 11179 compliant metamodel of Figure 4.8.

Data Elements and Value Domains

In this section the representational level of the metamodel is discussed. That is the registration of Data Elements and Value Domains related to the Data Element Concept 'Preservation Metadata for digital still images according to [NISOZ39.87: 2002]', which specifies the Conceptual Domain 'Preservation metadata for digital still images'. The documents of the ISO 11179 family of standards, as well as the related Technical Reports, contain a detailed and abstract overview of the way metadata should be registered. In this section the methods of the ISO/IEC 11179 family of standards and the concrete metadata elements of the [NISOZ39.87:2002] data dictionary are related to each other. For this, the metadata elements, as part of the [NISOZ39.87:2002], are considered to be Data Elements as defined by the ISO/IEC 11179 family of standards. The attributes that – according to the ISO/IEC 11179 family of standards – specify the definition, identification, representation and permissible values of Data Elements are analysed. This is done in two steps. First, the features of the metadata fields as described in the [NISOZ39.87:2002] document are analysed, followed by a projection of the principles derived from the Technical Reports [ISO20943-1:2003] and [ISO20943-3:2004] to establish the standardised attributes of the Data Elements.

The [NISOZ39.87:2002] data dictionary was published in 2002 as a 'Draft Standard for Trial Use'. On 31 December 2003 the draft period of 18 months ended.[155]

[155] In the year 2004 the status of the [NISOZ39.87:2002] did not change.

Table 4.7 Frequency of allowable data types in [NISOZ39.87:2002] data dictionary

Allowable data types in the [NISOZ39.87:2002] Data Dictionary	Definition	Number of occurrences in the [NISOZ39.87:2002] Data dictionary
DateTime	Representation of date and time	02
Enumerated type (restricted to external standard)	A string that may contain only one of a number of values as specified by an existing standard	19
Enumerated type (restricted to list)	A string that may contain only one of a number of values listed	21
Non-negative real	A real number where r >= 0	15
Positive integer	An integer where i > 0	16
Real	A real number where r may be 0	01
Reference	A single pointer to another object	04
String	One or more characters	32

According to the information in the [NISOZ39.87:2002] data dictionary at the end of the trial use period, the draft standard was returned to the development committee, along with any comments received for further action. The committee was to evaluate the comments and recommend further action. This information is relevant for the registration of the status of the Data Elements and has an impact on the attributes of the Data Elements.

Table 4.7 contains the short definition of the allowable data types as well as the frequency of occurrence of the data types in the [NISOZ39.87:2002] data dictionary. The data dictionary contains 111 Data Elements. Each Data Element (called 'TagName' in the document) can be identified with a unique name. For each Data Element the following seven attributes are registered in the data dictionary document: 'Definition'[156], 'Type', 'Required', 'Repeatable', 'Values (examples)', 'Notes or Usage Notes' and 'Use'. The following eight data types, as expression of the attribute 'Type', are used in the data dictionary: 'DateTime', 'Enumerated type (restricted to external standard)', 'Enumerated type (restricted to list)', 'Non-negative real', 'Positive integer', 'Real', 'Reference' and 'String'. Short definitions of the

[156] The attribute 'Definition' is given in Table 4.7.

data types are given in the [NISOZ39.87:2002] standard document. The committee that created the data dictionary had a good idea of the problem area to be covered by the data dictionary and used its expertise to draw up a list of metadata elements, classify the metadata elements under one of the four categories ('basic image parameters', 'image creation', 'imaging performance assessment and 'change history') and compile a list of attributes for each Data Element. An analysis of the allowable data types in the data dictionary reveals that (19 + 21 =) 40 of the 111 Data Elements in the [NISOZ39.87:2002] data dictionary are based on an existing enumerated domain of values (see: Table 4.7).[157]

Table 4.7 reveals that 19 enumerated Data Elements are based on an existing standard.[158] Also, several of the 21 enumerated Data Elements that are restricted to a list are based on existing standards. This implies that more than 30% of the Data Elements in the data dictionary are based on existing metadata elements. It can be concluded that the [NISOZ39.87:2002] set of metadata elements is based for a considerable part on existing metadata elements, making the establishment and usage of a metadata registry a relevant issue.

Following this review of the metadata elements of the [NISOZ39.87:2002] data dictionary, the theory concerning the standardised formulation of Data Elements according to the ISO/IEC 11179 family of standards will now be discussed.

A Data Element is a type of administered item (see Table 3.3). The [ISO11179-3:2003] standard document mentions 45 basic attributes common to all types of administered items. The basic attributes can be categorised as identifying, definitional, administrative and relational ([ISO11179-3:2003] pp. 55-59).[159] Additional attributes may be required when metadata items are used in a particular context.

The 45 basic attributes of administered items mentioned in the [ISO11179-3:2003] standard document play a part either in the Technical Report on the registration of Data Elements ([ISO20943-1:2003]) or in the Technical Report on the registration of Value Domains ([ISO20943-3:2004]). For example, the [ISO11179-3:2003] standard document distinguishes the attribute 'dimensionality'[160] as an attribute specific to Conceptual Domains. Technical Report [ISO20943-1:2003] on

[157] The total of occurrences of allowable data types is 110. As the data dictionary has 111 metadata elements this difference is explained by the fact that the metadata element 'previous image metadata' has the type '[retains previous data types]' ([NISOZ39.87:2002] p. 45).

[158] Of the 19 metadata elements that are based on a standard, 15 are based on the TIFF/EP standard ([ISO12234-2:2001]), three metadata elements are based on the DIG35 standard ([DIG00]) and one metadata element refers to the MIME extension.

[159] Figure 3.1 contains a model of the relation of an Administrated Item (such as a Data Element) with categories of attributes. The common facilities as mentioned in Figure 3.1 do not match the categories mentioned in the standard document ([ISO11179-3:2003]. The relational attributes are part of the 'classification' facility in Figure 3.1).

[160] The attribute 'Dimensionality' is defined as: 'The dimensionality for a concept. For example, length, mass, velocity, currency.' ([ISO11179-3:2003] p. 95). In the glossary of the standard document 'dimensionality' is defined as 'An expression of measurement without units' ([ISO11179-3:2003] p. 16).

Table 4.8 Number of mandatory, conditional and optional attributes of Data Elements as presented in Technical Report [ISO20943-1:2003]

	Category of attributes	M	C	O
1	Data Element definition	3	-	-
2	Permissible Values and Value Domain	7	4	4
3	Representation class	-	-	2
4	Data Element name and identifier	3	-	-
5	Other metadata attributes	4	-	2
6	Data Element Concept and Conceptual Domain	4	3	10
7	Data Element classification	-	1	1
8	Registration and administrative status information	2	-	-
9	Classification scheme for groups	4	3	1
	Total (58)	27	11	20

the registration of Data Elements mentions 'dimensionality' as a specific attribute but it is not mentioned in the list of metadata attribute names.[161] The Technical Report on Value Domains ([ISO20943-3:2004]), however, covers dimensionality as a feature of a Value Domain without referring to Data Elements. This example highlights the fact that the standard documents of the ISO/IEC 11179 family of standards and the Technical Reports that describe a set of procedures for the consistent registration of Data Elements lack interoperability. Technical Report [ISO20943-1:2003] is focused on the registration of Data Elements as a particular type of administered item and identifies a total of 58 attributes of the four core items of the metamodel that are either mandatory, conditional or optional. Table 4.8 contains an overview of the number of attributes per category.

A comparison of the [ISO11179-3:2003] standard document and the Technical Report that elaborates on the consistent registration of Data Elements ([ISO20943-1:2003]) reveals a discrepancy between the official standard document and the procedures as described in the Technical Report. The analysis of the NISO data dictionary on digital still images in relation to the ISO/IEC 11179 family of standards, aimed at the management of semantics of metadata, reveals that the application of the ISO/IEC 11179 standard requires the creation of a wide range of attributes that are often difficult to interpret. Instead of attempting to register all attributes of the ISO/IEC 11179 standard, the last part of this experiment consists of the expression of two metadata elements of the [NISOZ39.87:2002] data dictionary, based on the ideas as formulated by the ISO/IEC 11179 family of standards.

Two metadata elements from the [NISOZ39.87:2002] data dictionary are equipped with the essential attributes as formulated by the ISO/IEC 11179 family of standards. One Data Element has an enumerated Value Domain, while the other is connected to a non-enumerated Value Domain. Every Value Domain represents

[161] See Annex C 'Crosswalk of names in Technical Report to ISO/IEC 11179-3 metamodel' ([ISO20943-1:2003] pp. 95-97).

Table 4.9 Registration of a Data Element with an enumerated Value Domain

Attribute	Content of attribute
Data Element name:	Codes for compression schemes for digital image data
Data Element Concept context:	Basic digital image parameter
Data Element Concept name:	Compression scheme for digital image data
Data Element Concept definition:	The compression scheme used for a digital image
Conceptual Domain context:	Preservation metadata for digital still images
Conceptual Domain name:	Digital image compression schemes
Conceptual Domain definition:	Enumeration of digital image compression schemes
Value Domain name:	Digital image compression codes
Value Domain definition:	Codes for the compression scheme
Permissible values:	<1, Uncompressed>[162]
	<2, CCITT 1D>
	<3, CCITT Group 3>
	<4, CCITT Group 4>
	<5, LZW>
	<6, JPEG>
	<32773, Packbits>

two kinds of concepts: Data Element Concept (indirectly) and Conceptual Domain (directly). The Data Element Concept is the concept associated with a Data Element. Examples of Value Domains are given in Table 4.9 and Table 4.10.

The use of standardised codes ensures interoperability between metadata registries and application systems. The choice of codes is often arbitrary. In the example given in Table 4.10, permissible values are based on the TIFF version 6.0 standard. The number of permissible values may also be different. Each time a new compression scheme is added or subtracted, a new Value Domain, or Value Domain version, is created. Value Domains may be associated with many Data Element Concepts and, therefore, Data Elements. A list of permissible values (for instance, codes representing measurement units) can be used by several Data Elements. Value Domains do not have to be associated with a Data Element Concept at all. They may be managed independently. It may not make sense to manage a rapidly changing Value Domain in a registry. The many changes would require the formation, registration and administration of too many Value Domains. Alternatively, one could manage a rapidly changing enumerated Value Domain as a non-enumerated Value Domain. This has the advantage of maintaining all the metadata in the registry.

[162] In [ISO20943-3:2004] a permissible value is represented as an ordered pair delimited by angle brackets as follows: <value, value meaning> ([ISO20943-3:2004] p. 3).

Table 4.10 Registration of a data element with a non-enumerated value domain

Attribute	Content of attribute
Data Element name:	Medium of the analogue source material scanned to create a digital still image – text
Data Element Concept context:	Basic digital image parameter
Data Element Concept name:	Medium of the analogue source material scanned to create a digital still image
Data Element Concept definition:	The name of the general or specific physical nature of the analogue original that is scanned to create a digital still image
Conceptual Domain context:	Preservation metadata for digital still images
Conceptual Domain name:	Names for photographic techniques
Conceptual Domain definition:	Names representing photographic techniques
Value Domain name:	Textual English description of photographic technique
Value Domain definition:	Textual description of photographic technique
Non-enumerated Value Domain description:	English text
Examples:	Daguerreotype Reflection print Silver gelatine print Acme bronze 100 Chromagnetic film Microfilm

4.3.3 Conclusion

The practical examination of the registration of Data Elements according to the ISO/IEC 11179 family of standards as presented in the [ISO20943-1:2003] Technical Report reveals that the registration of a Data Element is very drawn out and can be very labour intensive. As not all attributes of Data Elements are compulsory, starting to register the most important attributes first can accelerate the registration process. An important feature of the Data Element registration process concerns the establishment of the registration and administration status. This makes it possible to start at a basic level and extend the quality of the Data Element in the course of time. For a consumer of a Data Element, the registration and administration status provides information on the relevance and importance of the Data Element.

The ISO/IEC 11179 family of standards is devoted to an accurate understanding of the semantics of Data Elements, and the metadata registry is important for the formalisation of the attributes of Data Elements. A registration authority, defined as an 'organisation responsible for maintaining a register' ([ISO11179-1:2004] p. 8), plays the central role in fixing the semantics of Data Elements. The registration authority has to bridge the gap between the abstract concepts of the ISO/IEC 11179 family of standards and the concrete assignment of Data Elements based on syntactical rules.

The experiment attempts to establish the semantics of the Data Elements, to represent the Data Elements and to register the Data Elements of the [NISOZ39.87:2002] standard in line with the ISO/IEC 11179 family of standards. The rationale behind this experiment is that a common understanding, reuse, harmonisation and management of the [NISOZ39.87:2002] Data Elements are all improved by using the ISO/IEC 11179 family of standards. The quality of the preservation metadata is an important facilitator for digital durability. In the second edition of the ISO/IEC 11179 family of standards, the metadata registry (MDR) became the central construct that enables the registration and identification of Data Elements. An MDR manages the semantics of Data Elements. The underlying model for an MDR is designed to capture all the basic components of the semantics of data, independent of any application or subject-matter area. Data Element descriptions have both semantic and representational components. The semantics are further divided into contextual and symbolic types.

As the current version of the ISO/IEC 11179 family of standards is a conceptual model and not intended to be a logical or physical data model for a computer system, the applicability of the ISO/IEC 11179 family of standards mainly has a value as a guideline for the formulation of unambiguous Data Elements. The application of ISO/IEC 11179 has to be seen in the light of the long tradition of methods for knowledge representation. Modelling requires interpretation of the properties and context of real-world objects. Often it seems that the results are too obvious. What is the difference between the data dictionary as a stand-alone text in natural language with explanatory remarks and the results of the top-down data element registration procedure? Complications can arise because people convey concepts through words (designations), and it is easy to confuse a concept with the designation used to represent it. The examples in the Technical Reports are very simple and result in very comprehensive schemes and tables. ISO/IEC 11179 is in the first instance a conceptual metamodel, not a physical implementation. Yet, some prescriptions are given. Some parts of the ISO/IEC 11179 family of standards do contradict each other. The attributes of Data Elements are not given in the same way in all documents. The experiment makes clear that the ISO/IEC 11179 family of standards is complex, is not stable and is difficult to apply.

5 Evaluation of the experiments

This chapter contains a careful judgement of the outcomes of the research up to now concerning the durability of digital surrogates of historical photographs. This evaluation chapter can be considered as a 'reality check' of the inductively formulated building blocks that enable long-term access to digital surrogates of historical photographs. The chapter consists of four parts. First, the main outcomes of the research so far are described briefly and put in a broader perspective. This analysis is the foundation for the next three sections of the chapter, each of which is devoted to an evaluation of one of the three experiments in chapter 4.

5.1 The main outcomes so far

For memory institutes, the durability of digital assets is important for two reasons. First, the creation of digital objects such as digital surrogates of historical photographs is expensive and digital archiving activities must prevent the loss of these valuable assets. Secondly, digital assets are used by memory institutes for a number of purposes, ranging from web access to the collection to the selling of printed reproductions of objects. Persistent access to digital surrogates of collection items is required to enable this multi-purpose usage. The concept of a 'use-neutral' digital master image is a synonym for a durable digital surrogate of a historical photograph. Also, the metaphor 'permanent pixels' applies to this concept.

Figure 5.1 gives an overview of the main outcomes of this research so far. A number of building blocks for digital preservation of digital surrogates of historical photographs have been formulated in an inductive way, by observing, studying and analysing relevant literature, projects and practices. Chapter 3 covers three inductively formulated premises relating to the creation of durable digital surrogates of historical photographs. This was realised by combining existing knowledge originating in a number of scientific disciplines that are not necessarily related to digital preservation. Chapter 4 consists of the examination of three hypotheses relevant for the research problem. The relation between chapter 3 and chapter 4 is illustrated in Figure 5.1.

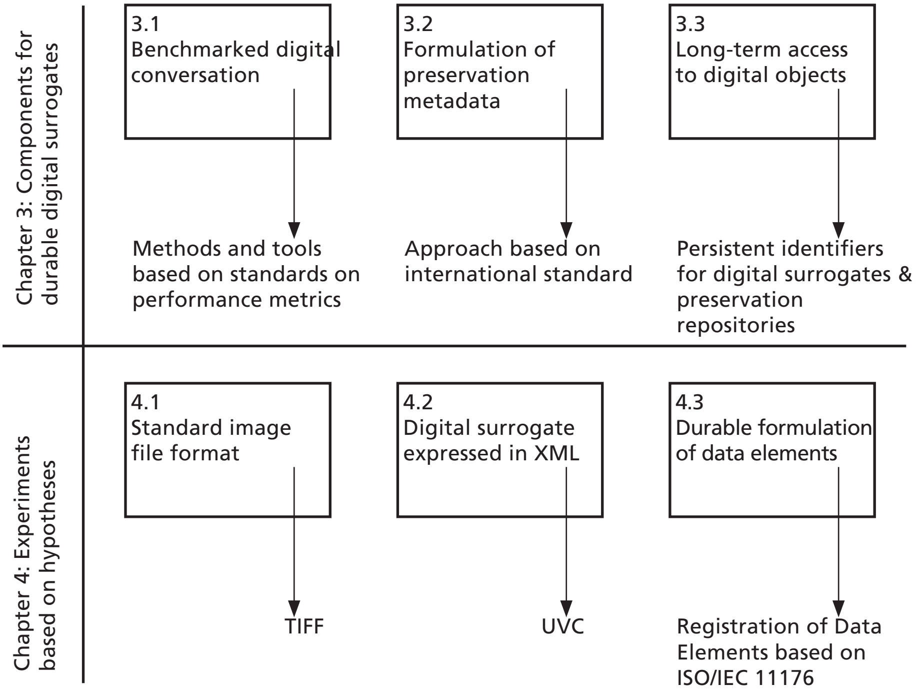

Chapter 3: Components for durable digital surrogates
3.1 Benchmarked digital conversation
3.2 Formulation of preservation metadata
3.3 Long-term access to digital objects
Methods and tools based on standards on performance metrics
Approach based on international standard
Persistent identifiers for digital surrogates & preservation repositories
Chapter 4: Experiments based on hypotheses
4.1 Standard image file format
4.2 Digital surrogate expressed in XML
4.3 Durable formulation of data elements
TIFF
UVC
Registration of Data Elements based on ISO/IEC 11176

Figure 5.1 Main outcomes of research so far

The first building block presented in chapter 3, relevant for the creation of durable digital surrogates of historical photographs, is a benchmarked digital conversion process, using methods and tools based on international standards on performance metrics. Predictive digital image quality is an important argument for the investments required for the creation and archiving of digital surrogates. The second building block concerns the importance of preservation metadata for providing long-term access to digital objects. An analysis of the activities of the International Organisation for Standardisation (ISO) results in the observation that the ISO/IEC 11179 family of standards potentially provides a framework for the unambiguous formulation of preservation metadata. The third building block covers durable storage of digital surrogates of historical photographs and results in the observation that persistent identifiers of digital objects as well as persistent repositories are important facets of this component.

Chapter 4 examines three inductively formulated hypotheses that enable the durability of digital surrogates of historical photographs. The first practical test in chapter 4 discussed the hypothesis that image file format standards are durable. The main outcome of this experiment is that the TIFF image file format is currently the most durable image file format and its usage has the lowest risk of future access to the digital images being hampered due to obsolescence of this file format.

The second experiment investigated the preservation of the bitstream that forms the digital image by using a non-proprietary data coding method. The standardised XML data format has a reputation as a durable data format. Three constructs – BSDL, UVC and XFlavor – use the XML data format for the coding of digital objects, such as digital images. Among the three constructs, the UVC approach contains the most complete solution for expressing the pixels of the digital image in XML as well as the content model that determines the meaning and position logic of the pixels. In addition, the UVC method is dedicated to playing a role in digital preservation whereas the other two constructs are related to the management of multimedia objects.

The third experiment tested how preservation metadata can be formulated unambiguously according to the emerging international ISO/IEC 11179 standard addressing the establishment of metadata registries. This experiment makes clear that data elements can be formulated in an unambiguous way, but the application of the ISO standard is complex and requires intensive training in order to understand the method for the unambiguous formulation of data elements.

5.1.1 Evaluation of the experiments

The three premises presented in chapter 3, which enhance the longevity of digital surrogates of historical photographs, and the three validated hypotheses of chapter 4 are constructed in an inductive way. Inductive reasoning implies using compelling arguments that are based on existing knowledge. Motivated assumptions and the current state of the art concerning aspects related to the long-term access of digital surrogates of historical photographs support the inductive strength of the preliminary conclusions of the research problem.

The evaluation of the experiments has also been carried out inductively, by giving arguments that either strengthen or contradict the outcomes of the experiments. More concretely, the following three issues have been clarified:

- Concerning the conclusion that the TIFF image file format enables long-term access to digital surrogates of historical photographs, recent developments regarding digital image formats have been studied that result in arguments that support or contradict this conclusion.
- Concerning the conclusion that the UVC is an optimal construct for encoding the pixels of digital surrogates of historical photographs in the durable XML data format, two evaluations have been carried out. First, the UVC as digital preservation method is placed in perspective with other strategies and, second, recent developments regarding the usage of the XML data format in relation to binary encoded objects have been studied.
- Concerning the standardised formulation and registration of preservation metadata (acting as an important indicator for the durability of the digital object that is documented with it), the metadata elements used by a number of information systems that provide access to digitised photographs of memory

institutes available on the Web were studied in order to find out to what extent the metadata elements can be incorporated in a metadata registry.

The next three sections of this chapter each cover one of the issues described above.

5.2 Emerging image file formats

As can be concluded from section 4.1, currently the standard image file format TIFF – more specifically Baseline TIFF version 6.0 – enables the long-term access of digital surrogates. Just as with any formal standard, two related factors are threatening the position of Baseline TIFF version 6.0: the appearance of competing new image file formats and the obsolescence of the existing TIFF image file format. Two new image file formats are gaining ground in the digital imaging community. These are the JPEG2000 digital image format and the RAW digital image format. Both image file formats are analysed with respect to their potential threat to the TIFF image file format as an optimal archival format for digital surrogates.

5.2.1 JPEG2000

Digital conversion and digital archiving applications by memory institutes are paying increasing attention to the JPEG2000 image file format. This is illustrated by the following two quotes: 'JPEG2000 offers significant advantages for digital archives. As an open international standard ... JPEG2000 is a superior format for the preservation of digital objects' ([JAN04] p. 148). 'Application areas for JPEG2000/Part 6[163] are document archiving by public authorities, libraries, ... or in general within industrial business archives... Archiving applications can efficiently deploy JPEG2000/Part 6 for historical and current documents' ([JUN04] p. 282). This focus on a new image file format will be examined in order to determine whether the TIFF image file format is indeed the most relevant format for encoding digital master images.

The ISO/IEC committee introduced the original JPEG image compression method in the late eighties. The JPEG specification is published as ISO/IEC standard [ISO10918-1:1994] and [ISO10918-2:1995]. JPEG compression is currently used on a large scale in areas where low data transmission speed and limited memory capacity are evident, such as Internet applications. The extension to the TIFF image file format supports the original JPEG compression, but the usage of this compression in digital master images is not recommended, as JPEG compression degrades the image quality and increases the risk of image file corruption. The original JPEG compression method is restricted by the maximal processing of 8 x 8 pixel blocks. Improved technical and algorithmic knowledge are the main forces behind the new JPEG2000 image compression method and format. The specifica-

[163] Part 6 of the JPEG2000 format 'Compound image file format' is directed towards the handling of mixed raster content, for instance printed text, graphics and images. See: <www.jpeg.org/jpeg2000/index.html> [cited 4 November 2004].

tions of the JPEG2000 image file format were published in 2001 as ISO standard [ISO15444-1:2004]. The JPEG2000 specification consists of 11 parts that will ultimately all be published as ISO standards.[164] The first part of the JPEG2000 specification, the core coding system, is the most relevant of the standard parts.

The original JPEG compression method uses a Fourier-related transformation, whereas the JPEG2000 compression method uses wavelet transformation and this is commonly regarded as a major improvement. Whether the JPEG2000 compression method yields higher-quality images compared with images compressed with the original JPEG compression method is still a subject of research (see, for instance, [STE03B]). The applied compression ratio turns out to be an important factor in the quality of the digital images. At high compression rates, the JPEG2000 method outperforms the initial JPEG compression method, but at medium and lower compression rates the JPEG compressed image is often assigned a higher quality ([STE03B] p. 580).

The image file compression algorithm is one of the specific features of the JPEG2000 file format. A number of other interesting properties of the JPEG2000 format are its high robustness in the presence of errors in the bitstream, its support of image-related metadata, facilities to provide protective image security features and the availability of a number of image resolutions included in one image file.

Despite the fact that the JPEG2000 image file format has a number of features that are lacking in the TIFF digital image file specification, in the near future the 'TIFF practice' (ref. [MUR04] p. 270) will prevail or at least exist alongside the usage of JPEG2000. The TIFF image file format was developed by the Desktop Publishing community and is, according to Murray, 'a previous technological innovation' ([MUR04] p. 270). The JPEG2000 standard originates in Imaging Science and Signal Processing and the effective implementation of JPEG2000 assumes familiarity with these disciplines. Murray argues that the technological language or 'digital library creole' currently used is based on concepts and terminology from the Desktop Publishing community. The creation and archiving of high-quality digital resources from previous technological innovations has yielded practices and practice-dependent points of view that make it difficult to appreciate and exploit the innovative characteristics of newer technologies. JPEG2000 has been well received and adopted by a small number of memory institutes and, as this number will grow, the perception and value of the TIFF digital image file format will change to the advantage of JPEG2000. But it is still too early to proclaim that the JPEG2000 format meets all digital preservation requirements.

5.2.2 RAW

Next to the JPEG2000 image file format the RAW image file format is gaining ground. High-end digital cameras create RAW image files. The RAW file format is

[164] The JPEG2000 specification and information on the parts of this specification can be found at: <http://www.jpeg.org/jpeg2000/index.html> [cited 3 November 2004]. Originally, the JPEG2000 specification consisted of 12 parts, but part 7 has been abandoned.

the uncompressed, unprocessed data file captured by the image sensor of the digital camera before any in-camera processing has been applied. The RAW image file format can be considered as the digital equivalent of the exposed but undeveloped film negative. Examples of common in-camera processing are sharpening, contrast and white balance manipulation. Instead of applying these operations, the settings are saved in a separate header associated with the RAW image data. The specification of the RAW image file format is not standardised. This results in the situation that each camera manufacturer has its own version. To mention some: 'Canon raw Format' (CRW), 'Nikon raw format' (NEF), 'Minolta raw format' (MRW), 'Olympus Raw Format' (ORF) are all proprietary versions of the RAW image file format.

Adobe have developed a digital image file format that is compatible with all versions of the RAW file format. This 'Digital Negative' (DNG) file format is a non-proprietary file format for storing camera raw files that can be used by a wide range of hardware and software vendors. The DNG format is an extension of the TIFF 6.0 format and is compatible with the TIFF-EP [ISO12234-2:2001] standard. The specifications of the DNG format can be found in [DNG04]. Adobe developed a plug-in that can be used by their image processing applications, such as Photoshop. The appearance of the RAW digital image file format is confirmation of the durability of the TIFF digital image file format, because the DNG format that is compatible with all versions of the RAW image file format is based on the TIFF digital image file specification.

Emerging digital image file formats do not contradict the premise that the TIFF digital image file format version 6.0 can currently be considered as the most durable digital image file format.

5.3 Bitstream preservation with the help of the XML data format

In this section, the role of the XML data format is evaluated as a construct for preserving digital surrogates of historical photographs. The hypothesis that the XML data format can be used to create 'permanent pixels' was examined in an experiment (see section 4.2) and resulted in the conclusion that the Universal Virtual Computer (UVC) approach is considered as the most appropriate of the usage of the XML data format for the encoding of durable digital images. In the first part of this section, the UVC approach is analysed more closely and, in the second part, initiatives originating in the XML community relating to the encoding of binary objects are described. These two cross-linked viewpoints between the digital preservation community (creating the UVC that uses the XML data format) and the XML development community (creating binary constructs to encode XML-encoded objects) serve as input for a deliberation on the role of the XML data format in a digital preservation framework for digital images.

5.3.1 Evaluation of the UVC approach

The UVC approach implies that for each file format and computer platform a UVC interpreter or virtual computer has to be developed. This is a labour-intensive and expensive activity. This leads to a high risk that the UVC approach will not result in a stable and reliable solution for long-term preservation of all types of digital surrogates.

The extensive user community of the TIFF digital image file format is also a factor contributing to bitstream preservation of digital surrogates with the UVC having a high risk of failure. Huge user communities of a specific file format will force the market to come up with backward compatibility if new formats are introduced. Two image file formats, DNG – which supports all variations of the RAW image file format – and JPEG2000 (see section 5.2), that were recently established, support this observation. The DNG format, developed by Adobe, is based on the TIFF digital image file format and the JPEG2000 digital image file format supports images that are compressed with the original JPEG compression method. This means that, if the DNG digital image file format is established in the future as the archival file format standard for digital surrogates, the master images currently created and stored as TIFF formatted images do not have to be converted. As the UVC method is not relevant for the long-term access of file formats with an extensive user community, the high development costs of a UVC will impede the development of a UVC interpreter for file formats that are used on a much smaller scale, but with a higher risk of becoming obsolete.

To use the durable XML data format for encoding the pixels of a digital surrogate image with the help of the UVC construct has practical drawbacks. This does not mean that the XML data format should not play a role in the longevity of digital surrogates. The METS 'wrapper' (see section 2.2.5) is an example of the role the XML data format can play in long-term access to digital objects.

An example of the function of the XML data format in the preservation of digital objects is the application of the ALTO data format, which is related to the METS specification. The ALTO data format (ALTO: 'Analysed Layout and Text Object') for digital versions of printed documents is an example of a file format using the XML data format that can be used to encode graphical data. The ALTO data format is a potential alternative to the UVC approach insofar as the usage of the XML data format is concerned. A bit is the smallest granular part of a digital object. The context of this bit is very important. Just as individual characters are meaningless, only words and other constructs are relevant. The ALTO data format was developed by the METAe project [MUH02].[165] This project addressed the development of application software for digital archives and libraries [STE03A]. The main goal of the application was the automatic recognition and extraction of metadata from printed material. The metadata engine developed by the project

[165] The website of the METAe project can be found at: <http://meta-e.uibk.ac.at/> [cited 12 May 2004].

creates an XML-formatted output file according to the METS standard. The output file acts as a Submission Information Package (SIP) according to the OAIS reference model [ISO14721:2003], ready for further processing and integration into a digital library. The project uses the DIG35 standard (ref. [DIG00]) for the creation of metadata of the digital images. Other data element sets used are Dublin Core metadata element set [ISO15836:2003] and MARC.[166] The METAe project designed the dedicated ALTO file format in order to be able to store detailed physical information of printed objects, such as font styles and layout characteristics. The ALTO file describes the logical structure and the physical structure of a digitised printed object such as a monograph or journal. The hierarchical layout structures, as well as information regarding the position of blocks and the text style, are assembled in an XML-formatted file. The ALTO data format makes it possible to encode and preserve almost all information gathered during, and available after, the automatic document analysis of the metadata engine.

5.3.2 XML binary working group

A drawback of using the XML data format instead of a proprietary binary format to encode multimedia objects such as digital images is the overhead resulting in an increased file size. XML tags consume a considerable amount of storage memory and bandwidth. Within the W3C an 'XML binary characterisation working-group' is active.[167] This working group has the task 'to gather information about use cases where the overhead of generating, parsing, transmitting, storing, or accessing XML-based data may be deemed too great for a particular application, characterising the properties that XML provides as well as those that are required by the use cases, and establishing objective, shared measurements to help judge whether XML 1.x and alternate (binary) encodings provide the required properties'.

The XML binary working group created a document in which relevant use cases are recorded in order to understand how the XML data format is currently being used and what optimisations are desirable and necessary.[168] Two use cases relate to the subject of this research:

- *Electronic documents*. A digital image is considered as an electronic document. In the use case, not only the static character of an electronic document but also the dynamic (for instance, animations) and interactive (for instance, form fields) character of electronic documents are taken into consideration.

[166] MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. See: <http://www.loc.gov/marc/> [cited 12 September 2004].

[167] The website of the W3C 'XML binary characterisation working group' can be found at: <http://www.w3.org/XML/Binary/> [cited 29 October 2004]. The working group was initiated after the organisation of the W3C workshop 'Binary interchange of XML information item sets', held in September 2003. A report on the workshop can be found at: <http://www.w3.org/2003/08/binary-interchange-workshop/Report.html> [cited 29 October 2004].

[168] See: <http://www.w3.org/TR/xbc-use-cases/> [cited 25 November 2004].

The XML data format is basically a syntax for marking documents, but is largely focused on textual data. The working group acknowledges that the XML data format fails to address the efficient embedding of binary encoded data, such as digital images. The TIFF digital image file format is mentioned as one of the alternatives for encoding binary objects, but as a shortcoming the lack of full support of the full range of required document features, such as dynamic and interactive content, is mentioned. These shortcomings are not relevant for the requirements for digital surrogates of historical photographs, resulting in the observation that the W3C binary working group considers the TIFF format a good alternative for the XML-formatted encoding of digital documents.

- *XML documents in persistent store*. This use case is based on the observation that documents are stored in a persistent storage facility. While it is not explicitly stated in the use case, it can be assumed that the documents are considered to be XML-formatted documents. The persistent store management system has to deal with many scenarios related to the size of the documents, the number of documents and the functions of the persistent storage system. For 'schema aware' documents (such as the XML-formatted images according to the UVC method), specific requirements of the persistent store are necessary, such as supporting schema evolutions. Regarding the possible alternatives, the working group is clear: '... persistent store vendors are required to handle XML to support multiple customer applications using XML. It would be absolutely inappropriate for persistent store vendors to offer their customers another format instead of XML'. This use case supports the assumption that the XML data format is suitable for the digital longevity of documents.[169]

The relevance of this evaluation of the role of the XML data format by the W3C XML binary working group for the longevity of digital surrogates is twofold. On the one hand, the working group considers binary encoding such as the TIFF digital image file format to be an appropriate alternative to the XML data format. On the other hand, the working group states that the XML data format has no alternative when it comes to the implementation of persistent storage.

5.4 Metadata of digital surrogates of historical photographs

Preservation metadata consists of a collection of metadata elements that are related to a digital object, intended to facilitate the long-term storage and access of that digital object. A number of metadata schemas have been developed that can be used to support the longevity of data objects. These metadata schemas have a wide range of functions, ranging from bibliographic description to documenting the technical features of data objects. Often, standard metadata schemas are extended with specific, local metadata elements, resulting in an application profile. An application profile consists of a collection of metadata elements that are the most appropriate for a specific organisation.

[169] The durability issues concerning the XML data format are covered in section 2.2.5

Ideally, standardised metadata schemas are the starting point for the compilation of an application profile. The usage of standardised metadata elements for the creation of preservation metadata has the advantage that users in the future can more easily interpret and understand the metadata elements. Standardised metadata elements can be considered as more durable than locally created metadata elements. Interoperability between collections that use similar metadata elements is also strengthened. As a consequence, standardised metadata schemas improve the durability of the objects they describe.

Metadata registries enable the efficient understanding, exchange and reuse of metadata elements. The importance of the unambiguous formulation of metadata elements is described in section 3.2 and whether the ISO/IEC 11179 family of standards can be used to create a metadata registry is examined in section 4.3. The 'reality check' on these metadata issues is the subject of this section.

5.4.1 Metadata elements used by memory institutes for access to digital surrogates of historical photographs

An indication of the quality of the metadata used by memory institutes can be gained by analysing the web interfaces to the digital image collections of memory institutes. The main function of a web interface is to provide access to items in a collection. Thus, it is foreseen that (provided they do exist) not all metadata elements that enable long-term access to the digital objects are visible to the users of the online access systems. Whether the reference image available on the Web, for instance in JPEG format, is a derivative of a high-quality digital master may not be apparent from the information at the web interface. Also, metadata elements that document the digitisation process may not be part of the web interface. Thus, the chances are that, more or less separate from the web interface, a more extensive digital preservation policy is established. Nevertheless, it can be assumed that the metadata elements used by the web interface of a digital collection of memory institutes give an indication of the way digital data objects are enriched by metadata elements and to what extent the longevity of the data objects is enabled by the metadata elements. A closer look at the web interface for digital collections of memory institutes will reveal to what extent the application of preservation metadata as described in this dissertation is used in practical situations.

Six web-accessible digital historical photograph collections have been evaluated. Two collections are part of the holdings of a library, two are part of a museum collection, and two digital historical photograph collections were made accessible by archival institutes. Each collection is based in a different country. By looking at six different countries, the scope of the observations is maximised and the risk of the systems being closely related is avoided. The systems used for the analysis are listed in Table 5.1.

Table 5.2 and Table 5.3 give an indication of the way memory institutes use metadata elements for the display of digital surrogates of historical photographs.

Table 5.1 Six web-based information systems of Memory institutes that give access to digital surrogates of historical photographs [cited July 2004]

Country	Name of system	URL of web interface	Type of Institute
New Zealand	TimeFrames, National Library of New Zealand	http://timeframes.natlib.govt.nz	Library
Austria	Bildarchiv, National Library of Austria	http://www.bildarchiv.at/	Library
United Kingdom	Photograph search, Imperial War Museum	http://www.iwmcollections.org.uk/qryPhotoImg.asp	Museum
France	Musée de la photographie (département de l'Essonne)	http://www.photographie.essonne.fr	Museum
Netherlands	Beeldbank Nationaal Archief	http://beeldbank.nationaalarchief.nl	Archive
Spain	Fototeca Sevilla	http://www.fototeca.us.es	Archive

Compared with web interfaces used by institutes that give access to collections of printed books, the systems that provide access to digital surrogates of historical photographs are very heterogeneous and require closer examination in order to make it possible to judge the quality and relevance of the collection. Not only the different languages used for the compilation of the labels of the metadata elements hinder a transparent indication of the value of the metadata elements, but also the various ways memory institutes document a digital surrogate of an historical photograph. This type of object is apparently less unambiguous, as may be expected. Most of the web interfaces contain background information on the meaning of the metadata elements as well as tips and hints for using the search system. Several metadata elements are connected to closed lists containing values that can be used in the input field. The access to distributed digital surrogates of historical photographs is heterogeneous and this affects the quality of the preservation metadata in a number of ways.

Based on inductive research, it became clear that the quality of preservation metadata is an important indicator and building block for the longevity of digital data. Chapter 3 contains an overview of the current state of the art regarding the compilation of preservation metadata. It can be concluded that, based on a global orientation of metadata elements dedicated to the access of digital historical photograph collections and accessible via web interfaces, it is difficult to determine the quality and relevance of the metadata elements. The clear and unambiguous formulation of metadata elements, as can be achieved by using the ISO/IEC 11179 family of standards, is not an issue for the memory institutes used for this orien-

Table 5.2 Search fields in web-based access systems to digital photograph collections

TimeFrames, National Library of New Zealand	*Bildarchiv, National Library of Austria*	*Photograph search, Imperial War Museum*	*Musée de la photographie*	*Beeldbank Nationaal Archief*	*Fototeca Sevilla*
Title	Suchbegriff(e)	Subject	Catégorie	Beschrijving	Nombre
Year	Person	Photographer	Sous catégorie	Dag/Maand/Jaar	Autor
Name	Schlagwort	Colour/B&W	Procédé	Periode vanaf/ tot en met	Localizacion
Subjects	Bildnummer	Period	Auteur	Trefwoord	Año entre ...y ...
Iwi/Hapu	Klassifizierung	Photograph Number	Auteur secondaire	Collectie	Tipo
Place	Medientyp	Collection Number	Titre	Fotograaf	Formato
Descriptive notes	Technik		Légende	Fotonummer	Serie
Image type	Institution		Sujet	Commentaar	Fotógrafo
Reference number	Jahr		Lieu de prise de vue		Fecha entre ... y ...
			Négatif á l'origine		Soporte
			Support		Sección
			Thème iconographique		
			Sous thème iconographique		
			Epoque		
			Date		
			Usage de destination		
			Usage connu		
			Recherche		

Table 5.3 Fields in result screen of web-based access systems to digital photograph collections

TimeFrames, National Library of New Zealand	Bildarchiv, National Library of Austria	Photograph search, Imperial War Museum	Musée de la photographie	Beeldbank Nationaal Archief	Fototeca Sevilla
Record title	Beschreibung	Photo No.	No. d'inventaire	Beschrijving	Objeto
Reference number	Inventar-nummer	Photographer	Catégorie	Datum	Fotógrafo
Display dates	Schlagworte	Collection title	Auteur	Trefwoorden	Autor
Quantity		Collection no.	Titre		Fecha
Restrictions		Description	Epoque ou date		Localizacion
Scope and content		Period	Procédé		
Historical names		Date	Description		
Places		Copyright	Support		
Search dates		Access	Marques et inscriptions		
		Colour / B&W	Dimensions hors tout		
		Type			

tation. It seems there is a gap between the theory for the creation of preservation metadata and the implementation in practice.

Based on an analysis of the practices of memory institutes (as illustrated by Table 5.2 and Table 5.3) for creating and using metadata elements, a number of conclusions can be drawn. Often, the attributes of the metadata elements are not defined; in other words they are not made explicit. Memory institutes do not mention the obvious Dublin Core metadata schema as a reference for the metadata elements of digital surrogates of historical photographs. This means that the metadata elements can be interpreted in a number of ways. In many situations the relation between the metadata element and the data object is not clear. It is not clear whether the metadata element refers to the scene visible on the image, the physical image or some other representation. A large gap exists between the theory of proper preservation metadata compilation and the reality, as can be seen by analysing the web interfaces of information systems that provide access to digital historical photograph collections.

5.4.2 Concluding: good practice for declaring metadata elements

Metadata elements are the key for access to digital resources. They provide essential information on the features and relevance of the digital resources. Carelessly and inaccurately created metadata elements will hamper the evaluation of the value of the resource by users. The declaration of metadata elements is not a trivial issue. Registries of metadata elements are important constructs for improving access to and interoperability between collections of digital objects, now and in the future. A shared set of principles underlying the construction of metadata registries is required. Also, agreement on common vocabulary for talking about the object of registries is of relevance.

The ISO/IEC 11179 family of standards provides unambiguous and precise descriptions of the nature, conditions of use and maintenance context of metadata elements, such that independent parties can understand, find and reuse them in other systems. The standard is not being used on a wide scale as yet. Currently, most metadata elements are declared in a variety of formats. Agreement on the formulation of metadata elements would move the web community one step closer to the integrated environment envisaged in the concept of the Semantic Web ([BAK02] p. 13).

Besides a standardised formulation of metadata elements, unique identification is also important. The W3C proposes making resources globally unique by using the URI specification. As metadata elements can be considered as resources, they can also be identified by a URI. In contrast to a URL, the URI does not have to resolve a particular document.

6 Conclusion

This research has addressed the durability of digital objects. The main reasons for the threatening loss of digital objects are the decay and obsolescence of the storage medium, the obscurity of the digital data format, and the insufficient quality or lack of metadata that is of major importance as the access key to the digital objects. Durability of digital objects is an issue that has emerged in the recent past and the three main general strategies for preventing decay of databases, digital documents, computer programs and other digital objects are technology preservation, technology emulation and digital information migration.

Digital objects are available in a wide range of functions and modes of complexity, and are supported by a variety of user communities and standard-setting organisations. Against the background of digital durability in general, the objective of this research has been to come up with concrete solutions for the durability of a specific type of digital object: the digital surrogate of an historical photograph. This case study approach has enabled the examination of standards, tools, methods and other digital durability building blocks in a concrete situation.

The creation, management and archiving of faithful digital surrogates of analogue original sources is a problem area specific to memory institutes, organisations that collect, select, store and provide access to all kind of artefacts created by society. Depending on the value and importance of the analogue original, for memory institutes the durability of digital surrogates is closely related to the connotation of a faithful, use-neutral and future-proof digital master file.

Often, digital preservation issues are discussed in general terms without clearly specifying the character and features of the digital objects that must be accessible in the long term. Contrary to this top-down approach, this study has focused on the durability of a specific category of digital object. Relevant strategies, standards and other building blocks related to digital longevity in general have been assessed, applied and evaluated in a detailed way. The concluding chapter of this dissertation consists of three parts. First, the research findings on building blocks for the durability of digital surrogates of historical photographs are presented. Also, a reflection on the research approach is given in the first part of this chapter. In the second part of the chapter, the outcomes of the research are extrapolated to the durability of digital objects in general. In the third part of the chapter, suggestions for further research are given.

6.1 Research findings

In the first chapter of this dissertation, the following research question was formulated: *How can benchmarked digital surrogates of historical photographs be preserved?* In order to answer this question, the following research objective was defined: *Identification and assessment of relevant building blocks that enable the creation, management and long-term access of benchmarked digital surrogates of historical photographs.* Research into digital preservation started about ten years ago and does not have a tradition or 'school of thought', making a number of scientific disciplines relevant for answering the research question. The inductive research strategy of this research is founded on the predictive strength of the arguments that are used to answer the research question by using correct reasoning and compelling arguments.

6.1.1 Main contributions

The main result of this research is the solid formulation and examination of building blocks for the longevity of digital images that are based on analogue historical photographs. Based on knowledge that originates from a number of sources, not necessarily related to digital preservation, in the third chapter three building blocks in the form of premises were introduced inductively and discussed. In chapter 4, three building blocks for digital durability were presented in the form of experiments that examine the validity of hypotheses upon which the durability of digital objects can be based. In the fifth chapter the experiments were evaluated.

Benchmarked capture, metadata formulation and durable access

Chapter 3 presented three premises that contribute to the durability of digital surrogates of historical photographs: the benchmarked digital capture of analogue originals, the unambiguous formulation of metadata, and durable access to digital objects using persistent identifiers and preservation repositories.

Benchmarked, faithful digital surrogates can be created by applying objectively measurable image quality settings based on international standards on digital imaging performance metrics that have become available in the recent past. This research examines international standards on digital imaging performance metrics specifically in relation to digital preservation issues. A benchmarked digital conversion process of an analogue photograph in essence implies the assessments of tone reproduction in the digital surrogate, detail and edge reproduction, image noise and colour reproduction by applying methods that are part of international standards.

The main function of metadata with regard to the durability of digital objects involves the attachment of documentation to digital objects in order to serve as the access key to them. Preservation metadata involves the registration of the essential characteristics of the digital object in order to retrieve the digital object in the future, to assess its value and to make the correct rendering of the object possible. This research concluded that a wide range of metadata schemas is potentially relevant for digital preservation and that in practice compilers and users of preservation metadata apply the 'mix and match' principle resulting in application profiles. Establishing the semantics of preservation metadata elements is of great importance in order for future generations of users to understand the meaning of the metadata elements and be able to process the digital object in a sensible way. Based on the assumption that the standardised formulation and application of metadata elements contributes to the quality of preservation metadata, thus improving the durability of the digital objects described with it, this research has evaluated a large number of ISO standards and given reasons for the role of the ISO/IEC 11179 family of standards as the point of departure for the unambiguous formulation of metadata elements.

In addition to a benchmarked conversion process, the unambiguous formulation of preservation metadata, the application of persistent identifiers and preservation repositories contribute to the longevity of digital surrogates of historical photographs. This research maps the current situation regarding the attribution of persistent, durable identifiers and the relevance of preservation repositories for long-term archiving of digital objects. Organisational support for persistent identifier resolution services turns out to be inevitable and of great importance for the durability of digital objects.

The concept of 'storage' is obviously much more pertinent to digital longevity than 'access'. However, a good digital preservation framework also takes access into consideration. This function forms a substantial part of the OAIS framework, as explained in chapter 2. An analysis of this function has shown that the harvesting interoperability principle fits best into a good digital longevity framework for digital visual sources. The OAI (Open Archives Initiative) Protocol for Metadata Harvesting (OAI-PMH) can be used to implement this requirement. The protocol supports any set of metadata elements, provided that these metadata descriptors are expressed in the XML data format and that the elements and attributes are stored in an XML Schema. It was concluded that the OIA-PMH is able to process ISO/IEC 11179 compliant metadata formats.

Experiments concerning image file format, bitstream encoding and metadata formulation

The examination of three hypotheses in chapter 4 in the form of experiments, and the evaluation of the experiments in chapter 5 clarify a number of issues and contribute to the results of this research. Regarding the data format of a digital image, that is the internal organisation of the data represented by the bits, the standardised image file format TIFF version 6.0 is a durable image file format. Despite the rapidly changing IT landscape, this digital image file format has been used on a large scale for more than ten years and new emerging image file format specifications, such as the EXIF and DNG specifications, are compatible with the TIFF digital image file format specification. In the longer term, the JPEG2000 digital image file specification has the potential to take over the role of the TIFF digital image file specification as the most durable digital image file format.

Based on the assumption that the XML data format enables the durable encoding of the bitstream that makes up a digital image, three methods that use the XML data format to enable the creation of application-independent multimedia objects have been evaluated. The UVC (Universal Virtual Computer) method has the most optimal approach regarding the conversion between a proprietary binary data representation, such as a digital image, and the recognised as durable XML data format. As the UVC method has its roots in the digital archiving community, this method has the potential to play an important role in durability infrastructures for digital images. The UVC method contains an emulator that understands the original format of a digital object as well as the processing specification of the digital object. The XML data format is used to express the logical view of a data object.

A drawback of the UVC method is the difficulty in making the logical view of the digital image explicit in an unambiguous way. On a more fundamental level, the XML data format can be questioned as an appropriate data format for the encoding of binary objects such as digital images. The W3C XML binary working group considers binary encoding for digital images, such as the TIFF image file format, a better alternative to the XML data format.

The third experiment concerns the standardised formulation of metadata based on the ISO/IEC 11179 family of standards. The relevance of the ISO/IEC 11179 family of standards for this research is that it is assumed that the standards might be used to establish the semantics of metadata elements, thus improving the quality of the metadata. Unambiguously formulated metadata elements will benefit the optimal usage of digital objects and contribute to their durability. A strong point of the ISO/IEC 11179 family of standards is the distinction between the conceptual level and the representational level of metadata schemas. The application of the ISO/IEC 11179 family of standards turned out to be a modelling exercise in which a number of approaches are possible. In theory, a conceptual level that takes all metadata schemas into consideration would result in Data Elements and Value Domains that can be used for the creation of preservation metadata for all conceivable types of digital objects, thus contributing to their durability. In practice, however, the conceptual level is closely related to a specific universe of discourse. In the experiment, for instance, this universe of discourse is the technical attributes of digital still images. Another obstacle to fixing the semantics of metadata using the ISO/IEC 11179 family of standards turned out to be the large number of attributes of administered items such as Data Elements and Value Domains that must be registered according to the ISO/IEC 11179 family of standards. The evaluation of this experiment revealed the lack of standardisation of metadata elements applied by memory institutes to make digital surrogates accessible on the Internet.

6.1.2 Reflection on the research approach

The research approach used in this dissertation is described in section 1.3 and consists of three elements: a research philosophy, a research strategy and research instruments.

Despite the fact that digital durability is increasingly recognised as a problem area and that a number of recognised global strategies have been formulated to prevent the loss of digital objects in the future, a research tradition does not exist. For this reason, in order to formulate a research approach for this study, first an inventory was complied of contributing scientific disciplines with related examples of research topics relevant for the preservation of digital surrogates of historical photographs. Next, methods for acquiring scientific knowledge in general were reviewed. This pluralistic view on science is interpretative by nature as it is based on subject-dependent perceptions and interpretations.

Inductive reasoning is the knowledge-gathering method that aligns with the nature of the research problem in this study. The induction problem, regarding the impossibility of deriving a universal conclusion based on a limited number of observations, inextricably related to the inductive research strategy, is analysed extensively. Despite the fact that some research at least claims to circumvent the induction problem, it must be concluded that the induction problem has not been solved. Thus, the predictive strength of the arguments that are used to answer the research question remains the main factor in determining the quality of the chosen research approach. This is the reason why accurate literature research and careful reasoning are the most important research instruments for this study. Also, a number of case studies involving active participation and implementation can be considered as research instruments. Relevant expert knowledge is the main factor in the predictive strength of the arguments used to address the research question.

Finally, the research strategy of this study consists of six activities that were carried out in the period 2000–2004 (see page 29). As information technology is changing fast, the risk existed that by the end of the research period the digital durability solutions proposed by the research would be unrealistic or outdated. The applied research approach, however, prevented the research arriving at conclusions that are misaligned with the current or future situation. Digital images provide a return far beyond the short term. A well-planned and executed digital imaging project has the potential to provide access to collections for decades to come.

6.2 Extrapolation

Taking the complete digital preservation landscape into consideration, this research can be considered as a case study. A main assumption on which this study is based is the observation that the detailed, thorough study of a specific well-defined and extensively described digital object will deliver outcomes that are relevant for the durability of a wide range of types of digital objects.

The outcomes of this research may not only be relevant for the durability of digital surrogates of historical photographs, but also for the durability of other digital objects, especially objects related to document management. In this concluding chapter, an attempt is made to extrapolate the findings of this research to the preservation of digital objects in general.

Before general remarks are made concerning the durability of digital data objects, first three issues are presented, that are specific to the durability of digital surrogates of historical photographs, that influence the required digital archiving activities. In the first place, the close relation between the analogue original and the digital surrogate justifies the importance of a benchmarked digital conversion process that strives to capture the significant details of the analogue original. Secondly, the relevance of a digital master file that can be used for multiple purposes without the need to re-scan the analogue original is typical for the durability of digital surrogates of analogue originals. Related to the first two issues is the frequently occurring vulnerability of the analogue original, which justifies the effort to create a high-quality digital surrogate. The three features described above are not necessarily relevant for the durability of born digital objects.

General issues regarding digital durability concern the applicability of one of the three main existing digital preservation strategies. These are technology preservation, migration of data and technology emulation. A second issue concerns the compilation and application of preservation metadata. The OAIS reference model, finally, provides the framework for the design of a solution for long-term storage of digital data objects.

Despite the fact that the high-level ambitions of the ISO/IEC 11179 family of standards have been only partly achieved, the standard is able to act as the framework for the creation of complete, comprehensive and consistent metadata elements required to support the longevity of digital objects. At the implementation level, it appears that the OAIS reference model is translated in different ways. A distinction can be seen between implementers of durable digital recordkeeping systems on the one hand and creators of preservation metadata frameworks for digital objects on the other. The former group consists mainly of representatives from the archival community who have a 'collection' perspective. The latter group can be found in libraries and museums where the 'item' perspective prevails. Due to the interpretation modes of the OAIS framework (or parts of the framework), a strict common ground on how to interpret this conceptual model can only be delineated within specific designated communities.

An archival data format must be utterly portable and self-describing, on the assumption that, apart from the transcription device, neither the software nor the hardware that wrote the data will be available when the data are read in the long term. Increasingly, digital images are digitally born – or 'born digital'. This is caused mainly by the growing usage of digital cameras. Long-term access to these digital images requires a proactive approach. This research makes clear that stakeholders must assess the importance and value of the images and act upon this. A risk analysis is part of this process. This study gives attention to a particular type of digital image in which the relation between features of the analogue original and the digital surrogate are important. It must be realised that this approach is based on a foundation with a long tradition in which the physical properties of information objects prevail over the intellectual content. The organisation of libraries is based on theories of the material characteristics and authors of books and other paper-based objects. This principle plays an important role in the authenticity issue of the object.

The reputation of the publisher as part of the bibliographic record of a book or magazine is one of the most important aspects in determining the relevance of the information object. As it will take some time before a common reference model for the value of digital objects is established, it is of importance to study the practice within the analogue world. Also, traditional libraries in the western world needed a considerable amount of time before a common framework was established and became part of the mental world of professionals in memory institutes.

This study makes clear that imaging science has produced methods and technologies that make it possible for memory institutes to create digital surrogates containing the same features, with regard to detail and spatial resolution, as the analogue original. These methods and technologies are increasingly being formalised in standards and these standards are used for the development of digitisation solutions that are becoming an integral part of digitisation equipment and software. It still requires some knowledge on the rationale of the settings and the way the targets should be used, but a digital capture device can be considered less and less as a 'black box' in which only the resolution output expressed in 'dpi' and the number of colours per pixel are the parameters to take into consideration. As digitisation is expensive, the conversion process must be based on a benchmarking phase. This research brings together the standards and methods that are required to achieve this benchmarked conversion. The objective quality issue can be of importance for the creation of durable digital surrogates of all kinds of analogue originals.

Application profiles are important building blocks for the creation of durable digital surrogates of digital objects. They contain all relevant metadata elements for a specific purpose for a designated community. It is stated that an application profile with descriptors to facilitate the longevity of digital surrogates of historical photographs should contain not only metadata elements that document the digital object but also descriptors that provide documentation of the original, analogue source. This is because characteristics of the original determine the specifications of the digital surrogate based on this original. Designated communities determine which descriptors are relevant for durable, 'use-neutral' digital masters that have a well-documented relationship with analogue sources. There is a risk that 'use-neutral' will turn into 'useless' if too many designated communities are addressed.

The creation of metadata is required to inform current and future users of the digital objects on the technical, administration, description, preservation and application aspects of the collection of digital objects. The monitoring of the medium is required to avoid sudden technical and physical obsolescence. The monitoring process often regulates the migration policy of data to a new storage medium. This 'digital cliff' can eventually lead to inaccessible or useless digital data. Documentation of the content, analogue source, purpose and applied technology that are related to the digital object plays an important role in the longevity of digital data.

Obsolescence of the data structure of the digital object is the main route to oblivion. This stresses the importance of standardised data formats. Interoperability requires common semantics. It is not the extensiveness of a metadata schema that is of importance but the quality of the metadata element formulation. Consistent terminology is important. Collection managers therefore need to work together with engineers and imaging scientists.

6.3 Further research

This research arrived at a point where a number of issues were selected as useful and relevant for the durability of digital surrogates of historical photographs. Suggestions for further research on durability of digital cultural heritage objects can be formulated in two directions. In the first place, on a micro level, the building blocks presented can be the subject of further research and on the other hand, on a macro level, further research can be carried out in the field of preservation of digital objects for memory institutes in general. Both perspectives are covered.

Regarding the building blocks for digital preservation of digital surrogates of historical photographs, the following directions for further research are relevant:

- A more detailed study should be carried out into the automatic digital capture of analogue originals based on recognised international image quality metrics standards. The present research made clear that a number of digital image quality assessment standards have recently become available. The application of these standards in order to arrive at benchmarked digital surrogates is still fairly complicated and labour intensive and further research should be aimed at facilitating the user-friendly application of the international quality standards.
- Further research addressing the assignment and management of persistent identifiers for digital surrogates as well as the implementation of preservation repositories would improve the durability of digital objects. This study

contains on overview of a number of persistent identifier schemes as well as architectures for preservation repositories. Further research is required in order to determine which solution fits best in a given situation. The research would imply more organisational issues than technical ones, as institutional support and commitment will be the key to the successful implementation of persistent identification of digital objects and storage of these objects in preservation repositories.

- The status of the TIFF digital image file standard must be monitored in order to determine the value of this standard for digital preservation purposes. Further research should be carried out that is aimed at monitoring the TIFF digital image file format and other digital image file formats as the most appropriate format for long-term storage. Current research on digital file format repositories and digital file format classifiers fits very well into this framework.
- The development of a UVC solution for the TIFF digital image file format is another suggestion for further research. Based on the ambitions of the emulation strategy of which the UVC is an implementation, it is feasible that an emulator for the TIFF image file format will become available in the near future. Position papers on the UVC approach refer to the emulation of complex processes, such as computer programs, making the creation of an emulator for the 'simple' TIFF digital image file format a fairly straightforward process. What requires thorough attention, however, is the formalisation of the logical view of the emulated digital object. This issue has already been mentioned in this research, but further research is required to solve it.
- Further research on the formulation, application and standardisation of preservation metadata should be carried out at a number of levels. As the revision of the ISO/IEC 11179 family of standards will be realised in the near future, further research is required to determine how a standardised metadata registry can be used to improve the durability of digital objects and how a metadata registry can be implemented according to the ISO/IEC 11179 family of standards.

Further research should also be carried out into the field of digital preservation for memory institutes in general. The DigiCULT Forum provides a number of points of departure for further research.[170] The report *The future digital heritage space* [GES04] summarises the results of an expedition into the possible future of digital heritage in the next 10 to 15 years. As 'persistent and perpetual access' is one of the themes discussed ([GES04] pp. 58-64), this publication can be used as a foundation for suggestions for further research into digital durability issues of digital objects.

[170] The website of the DigiCULT Forum can be found at: <http://www.digicult.info> [cited 19 November 2004]. The project is carried out as a support measure within the Information Society Technologies (IST) priority of the European Union's Fifth Framework Programme for Research and Technological Development.

The availability of digital heritage resources in an as yet unknown future seems more of an organisational issue than a technical one. Political, organisational, research and technological development issues are intertwined. Stable organisational, financial and expertise structures are important. An institutional preservation repository needs to be a service with continuity behind it. Institutions need to recognise they are making commitments for the long term. The diversity of digital objects makes problems of preservation and presentation very difficult and further research should be aimed at solving these problems.

The DigiCULT Forum states that not new but ongoing research has to be financed but it acknowledges that this is difficult to realise. Research must be better coordinated. Not projects, but institutes, have to be financed in order to arrive at a systematic preservation infrastructure. An increased understanding of the economics of long-term preservation is required and this is also an important area for further research. Research should also contribute to authoritative guidelines based on best practices such as risk management on digital preservation that will support the decision process in memory institutes.

Samenvatting

'Blijvende beeldpunten. Bouwstenen voor de duurzaamheid van digitale kopieën[171] van historische foto's

proefschrift van M.P.M. van Horik

Onderzoeksdomein

De toegang op lange termijn tot computerbestanden en computerprogramma's wordt bedreigd door de veroudering en kwetsbaarheid van digitale opslagmedia, het in onbruik raken van hardware en software en de opkomst van nieuwe digitale dataformaten. Er bestaat consensus over drie globale strategieën om het verlies van digitale objecten te voorkomen. Naast de strategie om in onbruik geraakte hardware en software in stand te houden waarmee verouderde programma's en bestanden kunnen worden verwerkt, bestaat de data-migratiestrategie, waarbij de digitale dataobjecten omgezet worden naar nieuwe gangbare formaten en opslagmedia en de strategie om verouderde hardware en software te emuleren op nieuwe computerplatforms. Emulatie impliceert dat verouderde systemen geïmiteerd worden op toekomstige generaties computersystemen.

In dit proefschrift wordt de digitale duurzaamheid van een specifiek digitaal dataobject onderzocht, namelijk digitale images die gebaseerd zijn op historische foto's. De motivatie om te kiezen voor een onderzoek naar de duurzaamheid van dit specifiek soort digitaal object heeft twee gronden. Op de eerste plaats maakt gerichte aandacht voor de duurzaamheid van een eenduidig digitaal object het mogelijk de relevantie van bestaande strategieën en bouwstenen die een rol spelen bij digitale duurzaamheid te toetsen in een concrete situatie. De behoefte van archieven, bibliotheken en musea om te weten hoe kwalitatief hoogwaardige digitale surrogaten van cultureel erfgoed objecten gemaakt kunnen worden en op welke

[171] In de Nederlandse samenvatting wordt 'surrogaat' als letterlijke vertaling van het Engelse 'surrogate' gebruikt. Deze begrippen hebben in beide talen een verschillende betekenis. In het Engels is een 'surrogate' een gelijkwaardige vervanging of substituut, terwijl in het Nederlands een 'surrogaat' een vervangingsmiddel van mindere kwaliteit is. In de Nederlandse samenvatting dient surrogaat in de Engelse betekenis te worden beschouwd. In de titel van de samenvatting wordt gebruik gemaakt van de term 'kopieën'.

wijze deze op lange termijn toegankelijk blijven vormt de tweede motivatie voor de keuze van de probleemstelling van dit proefschrift.

Het specifieke van een digitaal image dat gebaseerd is op een analoge historische foto is dat de relatie tussen het analoge origineel en het digitale surrogaat van essentieel belang is voor de kwaliteit en gebruiksmogelijkheden van het digitale surrogaat, nu en in de verre toekomst. Deze 'analoog origineel – digitaal surrogaat' relatie speelt een grote rol bij digitale conversieprojecten van instellingen die tastbaar cultureel erfgoed beheren. De bibliotheken, archieven en musea die het geheugen van de samenleving vormen, en daarom in dit proefschrift als 'memory institutes' worden aangeduid, streven naar de creatie van digitale surrogaten van analoge objecten die zo veel mogelijk de essentiële kenmerken van het origineel bevatten en nu en in de toekomst voor diverse doeleinden gebruikt kunnen worden. De creatie en archivering van deze 'use-neutral digital master files' is een kostbaar en arbeidsintensief proces.

Onderzoeksvraag en aanpak

De volgende onderzoeksvraag is geformuleerd: *'Hoe kunnen hoogwaardige digitale surrogaten van historische foto's toegankelijk blijven op lange termijn?'* Het doel van dit onderzoek kan als volgt geformuleerd worden: *'Het identificeren en evalueren van relevante bouwstenen die de creatie van en lange termijn toegang tot hoogwaardige digitale surrogaten van historische foto's mogelijk maken'.*

In dit onderzoek wordt een inductieve onderzoeksstrategie toegepast, waarbij zorgvuldig redeneren en het gebruik van correct geformuleerde argumentatie op basis van expertkennis centraal staat. Op inductieve wijze wordt getracht op basis van kennis uit het heden en verleden uitspraken te doen over de toekomstige toegang tot digitale objecten die momenteel of in het recente verleden gecreëerd zijn. Het inductieprobleem dat stelt dat het onmogelijk is een universele uitspraak te doen op basis van een beperkt aantal waarnemingen wordt zoveel mogelijk omzeild door gebruik te maken van conceptuele ruimten zoals deze door Gärdenfors worden geformuleerd. Een correct geformuleerde conceptuele ruimte kent geen concepten die gerelateerd zijn aan meer dan één dimensie of categorie, zoals kleur of tijd.[172] Anders gezegd betekent dit dat inductief redeneren bestaande kennis op een pragmatische wijze, rekening houdend met de context, activeert in praktische situaties.

Op basis van onderzoek naar bestaande praktijken en theorieën die relevant zijn voor het onderzoeksprobleem worden de concepten geformuleerd, die vervolgens het epistemologische vertrekpunt van het onderzoek vormen. Hoofdstuk twee van dit proefschrift heeft een inleidend karakter en bestaat uit twee onderdelen. De digitalisering van objecten door memory institutes wordt beschreven, gevolgd door een beschrijving van de stand van zaken ten aanzien van de duurzaam-

[172] Dit voorbeeld is gebaseerd op de 'Goodman Paradox', waarbij namen van kleuren gerelateerd zijn aan een tijdsperiode.

heid van digitale objecten. Verschillende wetenschappelijke disciplines leveren een bijdrage aan kennis en gereedschappen om duurzame hoogwaardige digitale surrogaten van historische foto's te maken en te archiveren. In hoofdstuk drie wordt deze kennis in kaart gebracht en gecombineerd. Drie vastgestelde aspecten die de duurzaamheid van een digitaal surrogaat vorm geven zijn: (1) de mate waarin de kenmerken van het analoge origineel op objectieve wijze zijn opgenomen in het digitale surrogaat, (2) de kwaliteit van de documentatie of metadata van zowel het digitale object als het analoge origineel en de relatie tussen het analoge origineel en het digitale surrogaat, en (3) de wijze waarop het digitale object toegankelijk is in een netwerkomgeving. Bovengenoemde drie aspecten worden scherper geformuleerd en onderbouwd met ondersteunende uitleg in de vorm van premissen, uitspraken die als feitelijk correct worden beschouwd.

Terwijl in hoofdstuk drie gebruik wordt gemaakt van bestaande inzichten die bij veel initiatieven op het gebied van digitale conversie en digitale duurzaamheid een rol spelen, bestaat hoofdstuk vier uit het gedetailleerd uitwerken van een aantal aspecten die van invloed zijn op de duurzaamheid van een specifiek digitaal object: het digitaal surrogaat van een historische foto. In hoofdstuk vier worden drie hypothesen die uit een analyse van het onderzoeksprobleem naar voren komen in de vorm van experimenten getoetst. De validiteit van de hypothesen wordt getoetst door deze in concrete operationele situaties toe te passen.

Het eerste experiment toetst de hypothese dat het gebruik van een gestandaardiseerd digitaal beeldformaat noodzakelijk is om digitale beelden op lange termijn toegankelijk te houden. De eigenschappen van gestandaardiseerde digitale beeldformaten worden geformaliseerd en er wordt vastgesteld welk digitaal beeldformaat aan deze eisen voldoet. In het tweede experiment staat het XML dataformaat centraal dat als duurzaam wordt beschouwd. Deze hypothese wordt getoetst door te onderzoeken of het XML dataformaat in staat is de individuele beeldpunten of pixels van digitale images te coderen. Het derde experiment, tenslotte, toetst de mate waarin het mogelijk is 'preservation metadata', noodzakelijk om het digitale image in de toekomst te interpreteren, eenduidig en helder te formuleren en te registreren. De onderliggende hypothese hier is de aanname dat eenduidig, expliciet geformuleerde metadata-elementen essentieel zijn voor de accurate toekomstige verwerking van de digitale data objecten die gedocumenteerd zijn met deze elementen. Verder wordt verondersteld dat de internationale standaard ISO/IEC 11179 geschikt is om deze eenduidige, expliciete formulering van metadata-elementen mogelijk te maken. In hoofdstuk vijf worden de experimenten geëvalueerd door, wederom op inductieve wijze, argumenten te formuleren die ofwel de resultaten van de experimenten bevestigen, dan wel ontkrachten. Hoofdstuk zes bevat de conclusie van dit proefschrift.

Verkregen resultaten

In hoofdstuk drie van dit proefschrift worden drie premissen gepresenteerd die van invloed zijn op de kwaliteit en duurzaamheid van een digitaal surrogaat van een analoge historische foto. Dit zijn achtereenvolgens de hoogwaardige digitale conversie van het origineel naar een digitaal surrogaat, het formuleren en toekennen van metadata aan de digitale objecten en de transparante gedistribueerde opslag en toegang tot de digitale objecten. Elk van de drie premissen wordt nader toegelicht.

In dit proefschrift is op basis van de ‘Image Quality Circle’ van Engeldrum – een raamwerk dat alle aspecten van beeldkwaliteit in beeld brengt - vastgesteld dat vooral de ‘Physical Image Parameters’ belangrijk zijn voor de kwaliteit van het digitale surrogaat en dat deze kwaliteit op een geijkt digitaliseringsapparaat gemeten kan worden. Tot voor kort was alleen een ‘imaging scientist’ in staat de prestaties van een digitaliseringsapparaat te meten en vast te stellen in hoeverre het apparaat de gewenste essentiële kenmerken van het origineel registreert en vastlegt in een digitaal beeldbestand. Inmiddels zijn er ISO en NISO standaarden, geijkte testkaarten en analysesoftware beschikbaar, zodat een conversie op basis van een ‘benchmark’ binnen het bereik komt van een instelling die historische fotocollecties op een kwalitatief hoogwaardige wijze wil digitaliseren. In dit proefschrift worden de standaarden waarmee hoogwaardige digitale images kunnen worden gemaakt beschreven en geanalyseerd.

Er is een groot aantal metadata-schema’s beschikbaar die van belang zijn voor de formulering van metadata van digitale objecten. In de praktijk wordt vaak het ‘mix and match’ principe toegepast waarin metadata-elementen afkomstig uit verschillende metadata-schema’s met elkaar worden gecombineerd, resulterend in een zogenaamd ‘applicatieprofiel’. Documentatie in de vorm van metadata is van groot belang voor de duurzaamheid van een digitaal object, omdat metadata essentiële informatie verschaft over de betekenis, waarde, formaat en andere kenmerken die relevant zijn voor het toekomstige gebruik van het digitale object. Een probleem is dat het moeilijk is de metadata-elementen die deel uitmaken van een metadata-schema met elkaar te vergelijken. Dit komt omdat de kenmerken van de data-elementen niet op een eenduidige manier zijn vastgelegd. Uit een analyse van beschikbare internationale standaarden kwam naar voren dat de ISO/IEC 11179 standaard het mogelijk maakt metadata-elementen eenduidig te formuleren. Het doel van deze standaard is om de formulering en het onderhoud van semantische beschrijvingen van data-elementen en waardedomeinen op een consistente en heldere manier te sturen. De standaard bestaat uit zes delen die allen voor het jaar 2000 de status van officiële ISO/IEC standaard hebben gekregen, maar in de periode daarna veranderd en aangepast zijn. In dit proefschrift wordt dit revisieproces nauwkeurig beschreven en geanalyseerd.

Transparante toegang via Internet tot digitale objecten, nu en op lange termijn, is een belangrijke indicator voor de duurzaamheid van de data. Digitale data die niet toegankelijk is heeft een grotere kans om verloren te gaan dan data die via Internet bereikbaar is, al dan niet afgeschermd door middel van een toegangscontrole om het intellectueel eigendom te beschermen. Deze premisse kan worden samengevat met 'access is preservation'. Authenticiteit en veiligheid zijn hierbij belangrijk. Authenticiteit om zeker te weten dat men toegang heeft tot het oorspronkelijke onveranderde object. Veiligheid behelst het voorkomen van onrechtmatig gebruik en het ongeautoriseerd manipuleren van de digitale objecten. Transparante toegang via Internet tot digitale objecten vereist de beschikbaarheid van bouwstenen in de vorm van standaarden, protocollen en software welke in dit proefschrift geanalyseerd worden. De mogelijkheden van het OAI-PMH protocol (Open Archives Initiative – Protocol for Metadata Harvesting) speelt een belangrijke rol bij de transparante gedistribueerde toegang via Internet tot digitale objecten. Een unieke, persistente 'identifier' is vereist om een digitaal object eenduidig te kunnen identificeren. Hiervoor zijn een aantal mogelijkheden beschikbaar, zoals het URI protocol, het OpenURL mechanisme, DOI (Digital Object Identifier) en het 'handle-system'.

De resultaten van hoofdstuk 4 bestaan uit de uitkomsten van drie experimenten welke elk een hypothese toetsen met behulp van beschikbare bouwstenen. Elk van de drie experimenten wordt nader beschreven. Het eerste experiment toetst de mate waarin een standaard bestandsformaat een waarborg is voor de langetermijn-toegang. De kenmerken van een gestandaardiseerd bestandsformaat voor digitale surrogaten van historische foto's worden vastgesteld en onderzocht wordt of er bestandsformaten zijn die voldoen aan deze kenmerken.

De zes inductief vastgestelde kenmerken van een standaard bestandsformaat zijn: (1) Het wordt gedurende een lange tijd door een grote gebruikersgroep gebruikt, (2) De specificaties zijn openbaar en worden ondersteund door een organisatie, zoals ISO. (3) Een breed scala aan systemen moet het bestandsformaat ondersteunen. Specifiek voor een grafisch bestandsformaat dat gebruikt kan worden voor de opslag van digitale surrogaten van historische foto zijn de volgende drie kenmerken. (4) Er mag geen datacompressie worden toegepast, omdat een gecomprimeerd bestand een grotere kans heeft om corrupt te raken en er bij datacompressie vaak sprake is van kwaliteitsverlies. (5) Het bestandformaat dient faciliteiten te bevatten om metadata op te slaan. Zonder documentatie is de functie en betekenis van een digitaal object immers moeilijk vast te stellen. (6) Het bestandsformaat dient alle essentiële kenmerken (zoals details en toonschaal) van het analoge origineel kunnen bevatten.

Op basis van de *Encyclopedia of Graphics File Formats*, gepubliceerd in 1994, kan worden vastgesteld dat een aantal bestandsformaten aan één of meerdere van de hierboven beschreven criteria voldoen. Alleen het TIFF formaat (Tagged Image File Format) voldoet aan alle criteria. Omdat er een aantal versies en extensies op het TIFF formaat bestaan, dient het bestandsformaat conform de gepubliceerde standaard gebruikt te worden.

Het tweede experiment toetst constructies waarmee de individuele beeldpunten of 'pixels' van een digitaal image op een duurzame manier kunnen worden opgeslagen. Omdat de pixels in de vorm van bits zijn opgeslagen in een digitaal beeldbestand, wordt gesproken van 'bitstream preservation'. De gestandaardiseerde opmaaktaal XML wordt beschouwd als duurzaam, omdat het XML dataformaat zelfbeschrijvend is, medium- en applicatie onafhankelijk, en gebaseerd is op gestandaardiseerde tekencodes die door alle computerplatforms kunnen worden geïnterpreteerd. Allereerst wordt de rol van het XML dataformaat gedetailleerd beschreven. Niet alleen de pixels zelf dienen in het XML dataformaat te worden weergegeven, maar ook de regels waaraan de ordening of structurering van de pixels in het beeldbestand moeten voldoen.

Er zijn drie methoden die het XML dataformaat gebruiken om de pixels van een digitaal image te coderen. Dit zijn BSDL (Bitstream Syntax Description Language), UVC (Universal Virtual Computer) en XFLAVOR (Formal language of Audio-Visual Object Representation).

BSDL is een op XML gebaseerde taal waarmee de structuur van een bitstream kan worden weergegeven. Elk formaat vereist een eigen BSDL representatie. De UVC-methode is gebaseerd op een 'universele virtuele computer', een 'interpreter' die instructies kan uitvoeren op elk hardware platform, nu en in de toekomst. Voor elk platform dient weliswaar eenmalig een interpreter te worden gebouwd, maar deze is dan wel in staat een bepaald dataformaat of computerprogramma waarvan de zogenaamde 'logical data view' in een XML structuur is vastgelegd, te begrijpen en te verwerken. Flavor is een formele taal bedoeld als methode om multimedia objecten te representeren zodat de objecten makkelijker gemanipuleerd kunnen worden. XFlavor is een extensie op Flavor waarbij het XML dataformaat wordt gebruikt om de multimedia-objecten te representeren.

De drie methoden worden op drie aspecten met elkaar vergeleken. (1) de manier waarop de methode een binair image kan converteren naar een digitaal document in het XML-dataformaat, (2) de manier waarop de methode in staat is de regels en de structuur van de pixels te formaliseren, en (3) de manier waarop (in de toekomst) de XML representatie weer omgezet kan worden in een binair- formaat. BSDL en XFlavor zijn het beste in staat om de structuur van een image te formaliseren in het XML dataformaat, maar het converteren van het binair-formaat naar een XML-formaat en het converteren van XML-formaat naar binair-formaat is problematisch. De UVC-methode is beter op dit punt, maar is weer beperkter in de mogelijkheden om de structuur van het image weer te geven.

Het derde experiment toetst de toepasbaarheid van de ISO/IEC 11179 standaard om 'preservation metadata' eenduidig te formuleren. Getracht wordt de metadata-elementen van een bestaande gestandaardiseerde NISO metadata-schema voor technische metadata van digitale images eenduidig uit te drukken met behulp van de ISO/IEC 11179 standaard. Op basis van beschikbare 'Technical Reports' gerelateerd aan de ISO/IEC 11179 standaard is de 'top-down' benadering betreffende de registratie van data-elementen toegepast. Hierbij wordt begonnen met het specificeren van het 'Conceptual Domain' en het 'Data Element Concept', die beide op abstracte wijze het domein beschrijven waarop de metadata-elementen betrekking hebben.[173] Naarmate het 'Conceptual Domain' algemener geformuleerd wordt neemt de mate waarin de data-elementen bruikbaar zijn voor andere domeinen toe. Daar staat tegenover dat de 'Data Elements' specifieker en eenduidiger worden naarmate het 'Conceptual Domain' concreter wordt. Er worden drie scenario's geschetst volgens welke de 'top-down' registratie van Data Elements (op het gebied van technische metadata van digitale images) uitgevoerd kan worden. Hierbij blijkt dat het scenario waarbij het 'Conceptual Domain' het meest specifiek is en gedefinieerd is als 'Preservation metadata for digital still images'[174] het beste aansluit bij de structuur en inhoud van de NISO standaard op het gebied van technische metadata voor digitale images. Dit scenario wordt verder uitgewerkt, waarbij blijkt dat een bestaand XML-Schema, het MIX XML-Schema, uitstekend kan dienen als datamodel voor de metadata-elementen van de NISO-standaard. Ten slotte zijn de 'Data Elements' en 'Value Domains' gedocumenteerd conform de ISO/IEC 11179 standaard.

Het experiment maakt duidelijk dat de registratie van Data Elements conform de ISO/IEC 11179 standaard niet eenvoudig is. Het registreren van de kenmerken van een Data Element is een zeer arbeidsintensief proces. De Technical Reports die als doel hebben te beschrijven hoe de abstracte ISO/IEC 11179 standaard in de praktijk dient te worden toegepast bevatten tegenstrijdigheden en dit maakt het gebruik van de standaard moeilijker. Het doel van de ISO/IEC 11179 standaard is het eenduidig vastleggen van de semantiek van metadata in de vorm van een 'Registry', maar zonder subjectieve interpretatie van de standaard is het niet mogelijk deze toe te passen. De ISO/IEC 11179 standaard kan zeer zeker dienen als richtlijn om metadata voor digitale objecten te creëren, maar de standaard is niet concreet genoeg om zonder problemen de kenmerken van metadata te formaliseren, waardoor het hergebruik en de uitwisseling van metadata-elementen vergemakkelijkt zou kunnen worden. Er wordt actief gewerkt aan de standaard, waardoor het mogelijk is dat in de toekomst de standaard makkelijker toepasbaar zal worden, bijvoorbeeld door de beschikbaarheid van ondersteunende software.

Concluderend kan worden gesteld dat dit proefschrift aantoont dat er bouwstenen beschikbaar zijn om duurzame digitale surrogaten van historische foto's te maken. Digitalisering van de analoge originelen dient plaats te vinden op basis van een beoordeling van de essentiële kenmerken van het origineel, zoals de fysieke verschijningsvorm, de details en de toonschaal. IJking van het digitaliseringsapparaat op basis van testkaarten en speciale analysetechnieken garandeert een digitaal surrogaat met een objectief toetsbaar kwaliteitsniveau. Metadata is van groot belang om toekomstige interpretatie van het digitale object mogelijk te maken. Er is een aantal al dan niet gestandaardiseerde sets van data- elementen beschik-

[173] Voor de definitie van 'Data Element', 'Value Domain', Conceptual Domain' en 'Data Element Concept', zie: Table 3.3.
[174] Zie Table 4.6.

baar, maar een kritische analyse van de waarde en functie van de data-elementen is essentieel. Door waar mogelijk is het XML dataformaat te gebruiken neemt de kans toe dat ook in de verre toekomst de gegevens kunnen worden benaderd. Of deze gegevens ook kunnen worden begrepen en verwerkt hangt af van de mate waarin de structuur van de XML-bestanden zelfbeschrijvend is en of het bestand waarin de syntax van de XML-bestanden is vastgelegd ook in de toekomst begrepen wordt.

References

Literature[175]

[ABR03] Abrams, S. and D. Seaman, 'Towards a global digital format registry'. [online] In: *Proceedings of 69th IFLA congress*, Berlin, 2003 [cited 23 March 2004]. PDF format. Available from World Wide Web: <http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf>.

[ABR04] Abrams, S. and D. Seaman, 'Global digital format registry'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 83-87.

[AHM01] Ahmed, K. *et al.*, *Professional XML Meta Data,* Birmingham (Wrox Press), 2001.

[ALE00] Alemneh, D., S. Hastings and C. Hartman, 'A metadata approach to preservation of digital resources: the university of North Texas Libraries' experience'. [online] In: *First Monday,* vol. 7 (2002) [cited 23 March 2004]. HTML format. Available from World Wide Web: <http://firstmonday.org/issues/issue7_8/alemneh/index.html>.

[AMI02] Amielh, M and S. Devillers, 'Bitstream syntax description language: Application of XML-Schema to multimedia content adaption' [online]. In: *Proceedings of the WWW2002: the 11th International World Web Conference*, Honolulu, 2002 [cited 24 March 2004]. HTML format. Available from World Wide Web: <http://www2002.org/CDROM/alternate/334/>.

[AND92] Anderson, S., 'The future of the present. The ESRC data archive as a resource centre of the future'. In: *History and Computing*, vol. 4 (1992), pp. 191-196.

[ANT04] Antoniou, G. and F. van Harmelen, *A semantic web primer,* Cambridge, Mass. (MIT Press), 2004.

[ARM01] Arms, C., 'Learning from American Memory: Opportunities and chal-

[175] The reference to electronic documents is stated according to ISO 690-0:1997 *Information and documentation – Bibliographic references – Part 2: Electronic documents or parts thereof.* International Organisation for Standardisation.

lenges ahead'. In: *Proceedings of Kyoto International conference on digital libraries: research and practice*, Los Alamitos (IEEE Computer Society), 2001, pp. 46-48.

[AUB98] Aubenas, S., 'The photograph in print. Multiplication and stability of the image'. In: M. Frizot (editor), *A new history of photography*. Köln (Könemann Verlagsgesellschaft),1998, pp. 225-231.

[BAK00] Baker, Th., 'A grammar of Dublin Core' [online]. In: *D-Lib Magazine*, vol. 6 (2000) [cited 12 February 2004]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib /october00/baker/10baker.html>.

[BAK01] Baker, Th. *et al.*, 'What terms does your metadata use? Application profiles as machine-understandable narratives' [online]. In: *Journal of digital information*, issue 2 (2001) [cited 23 April 2004]. PDF format. Available from World Wide Web: <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Baker/baker-final.pdf>.

[BAK02] Baker, Th. *et al.*, *Principles of Metadata registries*. White paper by the DELOS Working Group on Registries [online] 2002 [cited 23 April 2004]. PDF format. Available from World Wide Web: <http://www.lu.lv/szf/BZIZN/DELOS/images/Publ_WPRegistries_A.pdf>.

[BAK03] Baker, Th. and M. Dekkers, 'Identifying metadata elements with URIs. The CORES resolution' [online]. In: *D-Lib Magazine*, vol. 9 (2003) [cited 23 March 2004]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/july03/baker/07baker.html>.

[BAT03] Battiata, M., 'Buried treasure. Why has Bill Gates stashed millions of the greatest images of the 20th century under a mountain in Pennsylvania?' In: *Washington Post Magazine*, Sunday May 18 2003, p. W14.

[BEN02] *Benchmark for faithful digital reproductions of monographs and serials.* Report by the Digital Library Federation Benchmark Working Group (2001-2002) [online] [cited 12 June 2004]. HTML format. Available from World Wide Web: <http://purl.oclc.org/DLF /benchrepro212>.

[BER69] Bertels, K. and D. Nauta, *Inleiding tot het modelbegrip*. Bussum (De Haan), 1969.

[BER00] Berns, R., *Billmeyer and Salzman principles of color technology*. New York (Wiley), 2000.

[BER02] Berns, R., 'Sneaking scientific validity into imaging tools for the masses'. In: *Proceedings of the first European IS&T Conference on color graphics and vision*, (Society for imaging science and technology), 2002, pp. 1-2.

[BES96] Besser, H. and J. Trant, *Introduction to imaging. Issues in constructing an image database.* Santa Monica (Getty Research Institute), 1996. Also available in HTML version from World Wide Web: <http://www.getty.edu/research/conducting_research/ standards/introimages/> [cited 5 October 2004].

[BEY03] Beyers, F., *Care and handling of CDs and DVDs. A guide for librarians and archivists.* [online] Washington D.C. (CLIR Reports), 2003 [cited 28 September 2004]. PDF format. Available from World Wide Web: <http://www.clir.org/pubs/reports/pub121/pub121.pdf>.

[BIK93] Bikson, T.K. and E.J. Frinking, *Preserving the present* (Het heden ont-houden). Den Haag (SDU), 1993.

[BOO96] Boom, M. and H. Rooseboom, *Een nieuwe kunst. Fotografie in de 19e eeuw. A new art. Photography in the 19th century*. Amsterdam (Rijksmuseum / Van Gogh museum), 1996.

[BOO04] Boonstra, O., L. Breure and P. Doorn, *Past, present and future of historical information science.* Amsterdam (NIWI-KNAW), 2004.

[BOR95] Born, G., *The file formats handbook*. London/Boston (International Thomson Computer Press), 1995.

[BRA98] Bradley, N., *The XML companion.* Harlow (Addison Wesley), 1998.

[BRO95] Browne, C.W. and B.J. Sheperd, *Graphics file formats. Reference and guide*. Greenwich (Manning), 1995.

[BUR01] Burns, P. and D. Williams, 'Distilling noise sources for digital capture devices'. In: *Proceedings of IS&T Image processing, Image Quality, Image Capture (PICS) Conference*, Montreal, Quebec (Society for imaging science and technology), 2001, pp. 132-136.

[CHA99] Chalmers, A., *Wat heet wetenschap?* Amsterdam (Boom), 1999. First edition: 1976 (What is this thing called science?).

[CHA03] Chapman, S., 'Counting the costs of digital preservation: is repository storage affordable?' [online]. In: *Journal of digital information,* vol. 5 (2003). [cited 24 April 2004]. PDF format. Available from World Wide Web: <http://jodi.ecs.soton.ac.uk/ Articles/v04/i02/Chapman/chapman-final.pdf>.

[CHA04] Chapman, S. and L. Abrams, 'Steering resources to safe-harbor repositories: the need for reliable, accurate and affordable ingest services'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 98-102.

[CHI01] Chilvers, A., 'The super-metadata framework for managing long-term access to digital data objects'. In: *Journal of documentation*, vol. 58 (2002), pp. 146-174.

[CHI03] 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)* [online] 2003 [cited 24 July 2004]. PDF format. Available from World Wide Web: <http://www.delosnsf-imagewg.unifi.it/assets/WG_final_report.pdf>

[CHO03] Choi, Y. and E. Rasmussen, 'Searching for images: The analysis of users' queries for image retieval in american history'. In: *Journal of the Ameri-*

can society for information science and technology, vol. 54 (2003), pp. 498-511.

[COE83] Coe, B. and M. Haworth-Booth, *A guide to early photographic processes*. London (Victoria & Albert Museum), 1983.

[CRO02] Crowe, R. *The Case for Institutional Repositories: A SPARC position paper* [online] Washington D.C. (SPARC), 2002 [cited 23 October 2004]. PDF format. Available from the World Wide Web <http://www.arl.org/sparc/IR/IR_Final_Release_102.pdf >.

[CRO03] Crofts, N., M. Doerr and T. Gill, 'The CIDOC Conceptual Reference Model. A standard for communicating cultural contents' [online]. In: *Cultivate Interactive,* issue 9, (2003) [cited 3 April 2004]. HTML format. Available from World Wide Web: <http://www.cultivate-int.org/issue9/chios>.

[DAC03] Daconta, M., L. Obrst and K. Smith, *The semantic web: a guide to the future of XML, web services, and knowledge management*. Indianapolis (Wiley), 2003.

[DAL04] Dale, R. and G. Waibel, 'Capturing technical metadata for digital still images' [online]. In: *RLG Diginews*, vol. 8 (2004) [cited 20 October 2004]. HTML format. Available from World Wide Web: <http://www.rlg.org/en/page.php?Page_ID=20462#article1>.

[DAY01] Day, M., 'Metadata for digital preservation: a review of recent developments'. In: *Proceedings of 5th European Conference for Digital Libraries*. Darmstadt (Springer), 2001, pp. 161-172.

[DIE02] Diessen, R. van, and J. Steenbakkers, *The long-term preservation study of the DNEP project. An overview of the results*, The Hague (Koninklijke Bibliotheek), 2002. Also online available in PDF format from World Wide Web: <http://www.kb.nl/hrd/dd/dd_onderzoek /reports/1-overview.pdf> [cited 7 June 2004].

[DIN01] Ding, Y., 'A review of ontologies with the Semantic Web in view'. In: *Journal of information science*, vol. 6 (2001), pp. 377-384.

[DOL00] Dollar, Ch., *Authentic electronic records: Strategies for long-term access*. Chicago (Cohasset Associates), 2000.

[DON03] Dondorp, F. and K. van der Meer, 'Design criteria for digital repositories' [online]. In: *Proceedings Conferentie Informatiewetenschap,* Eindhoven (TU Eindhoven, Werkgroep Informatiewetenschap, Onderzoeksschool SIKS), 2003 [cited 28 December 2004]. HTML format. Available from World Wide Web: <http://wwwis.win.tue.nl/infwet03/proceedings/4/>.

[DOO89] Doorn, P. and N. van Hall, *Nederlands Historisch Data Archief. Eindverslag van een verkennend onderzoek*, Amsterdam (SWIDOC/Steinmetzarchief – VGI), 1989.

[DOO90] Doorn, P., R. van Horik and J. Touwen, *Nederlands Historisch Data Ar-*

chief I. Eindverslag van een pilot project, Amsterdam (SWIDOC/Steinmetzarchief – VGI), 1990.

[DOO92] Doorn, P. and H. Matthezing, 'After the flood: Archiving electronic records in the Netherlands'. In: *History and Computing*, vol. 4 (1992), pp. 197-200.

[DOO96] Doorn, P., 'Digital archives'. In: C. Mullings (editor) *New Technologies for the Humanities*, London (Saur), 1996, pp. 357-379.

[DOO04] Doorn, P., I. Garskova and H. Tjalsma (editors), *Archives in cyberspace. Electronic records in East and West*, Moscow (Moscow University Press), 2004.

[DOR92] Dooren, W. van, *Vragenderwijs. Elementair overzicht van de systematische filosofie,* Assen/Maastricht (Van Gorcum), 1992.

[DUR01] Dürr, E. and K. van der Meer, *Emulation and conversion: Organisational overview – way of working, costs, methods. Report at the E-Archive project; v1.2* [online] 2001 [cited 5 October 2004]. PDF format. Available from World Wide Web: <http://www.library.tudelft.nl/e-archive/Documenten/Resultaten/roquade2.pdf>.

[DUR04] Duranti, L. and J. Blanchette, 'The authenticity of electronic records: the InterPARES approach' In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 215-220.

[DUV02] Duval, E. *et al.*, 'Metadata principles and practicalities' [online]. In: *D-Lib Magazine*, vol. 8 (2002) [cited 23 February 2004]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/april02/weibel/04weibel.html>.

[ELE03] Eleftheriadis, A. and D. Hong, 'Flavor: a language for media representation'. In: B. Fuhrt and O. Marques *Handbook of video databases.* Boca Raton (CRC Press), pp. 69-96, 2003.

[ENG99] Engeldrum, P. 'Image quality modeling: Where are we?' In: *Proceedings of IS&T Image processing, Image Quality, Image Capture (PICS) Conference*, Savannah, Georgia (Society for imaging science and technology), 1999, pp. 251-255.

[ENG04A] Engeldrum, P., 'A short image quality model taxonomy'. In: *Journal of imaging science and technology,* vol. 48 (2004), pp. 160-165.

[ENG04B] Engeldrum, P. 'A theory of image quality: The image quality circle'. In: *Journal of imaging science and technology*, vol. 48 (2004), pp. 447-457.

[EST96] Ester, M., *Digital image collections: Issues and practice. Washington D.C.* (the Commission on preservation and access), 1996.

[FEN03] Fensel, D., F. van Harmelen and I. Horrocks, 'OIL and DAML + OIL: Ontology languages for the semantic web'. In: J. Davies, D. Fensel and F. van Harmelen (editors), *Towards the semantic web. Ontology-driven knowledge management*, Chichester (Wiley), 2003, pp. 11-31.

[FIN02] *Findings on the preservation of authentic electronic records*. Final report to the National Historical Publications and Records Commission (Grants # 99-073 and # 2001-005) (US-InterPARES Project, International Research on Permanent Authentic Records in Electronic Systems), September 2002.

[FLE92] Fleischhauer, C. and R. Erway, *Reproduction-quality issues in a digital-library system. Observations on the reproduction of various library and archival material formats for access and preservation*. An American Memory White Paper [online] (1992). [cited 24 November 2003]. ASCII format. Available from the Internet: <ftp://rs7.loc.gov/pub/american.memory/white.papers/reprod.txt>.

[FOL96] Foley, J., A. van Dam, S. Feiner and J. Hughes, *Computer graphics. Principles and practice*. Reading (Addison-Wesley), 1996.

[FRE97A] Frey, F., 'Digitization of photograph collections'. In: *Proceedings of the IS&T 50th annual conference*. Cambridge, Massachusetts (Society for imaging science and technology), 1997, pp. 597-599.

[FRE97B] Frey, F. 'Digitize to preserve – Photographic collections facing the next millennium'. In: *Proceedings of the IS&T 50th annual conference*, Cambridge, Massachusetts (Society for imaging science and technology), 1997, pp. 713-715.

[FRE97C] Frey, F. and S. Süsstrunk, 'Color issues to consider in pictorial image data bases'. In: *Proceedings of IS&T fifth color imaging conference*, Scottsdale, AZ (Society for imaging science and technology), 1997, pp. 112-115.

[FRE99] Frey, F. and J. Reilly, *Digital imaging for photographic collections. Foundations for technical standards.* Rochester (Image Permanence Institute, Rochester Institute of Technology), 1999.

[FRE00A] Frey, F., 'Measuring quality of digital masters' [online]. In: *Guides to quality in visual resource imaging* (DLF, RLG, CLIR) 2000 [cited 19 September 2003]. HTML format. Available from World Wide Web: < http://lyra2.rlg.org/visguides/visguide4.html >.

[FRE00B] Frey, F., 'File formats for digital masters' [online]. In: *Guides to quality in visual resource imaging* (DLF, RLG, CLIR) 2000 [cited 19 September 2003]. HTML format. Available from World Wide Web: <http://lyra2.rlg.org/visguides/visguide5.html>.

[FRE00C] Frey, F. and S. Süsstrunk, 'Digital photography – How long will it last?' In: *Proceedings of IEEE ISCAS 2000*, vol. 5 (2000) pp. 113-116.

[FRE01] Frey, F., 'Developing specifications for archival digital still images'. In: *Proceedings of IS&T Image processing, Image Quality, Image Capture (PICS) Conference*, Montreal, Quebec (Society for imaging science and technology), 2001, pp. 166-171.

[FRI98] Frizot, M. (editor), *A new history of photography.* Köln (Könemann Verlagsgesellschaft), 1998. (English language version of French edition published in 1994).

[FUA98] Fuad-Luke, A., *Digital photography. How to capture, manipulate and output images*. London (Guardian), 1998.

[FUN01] Fung, K., *XSLT. Working with XML and HTML.* Boston (Addison-Wesley), 2001.

[GAN99] Gann, R.G., *Desktop scanners. Image quality evaluation.* Upper Saddle River (Prentice Hall), 1999.

[GAR98] Gardner, S., 'Kant'. In: A.C. Grayling (editor), *Philosophy 2.* Oxford (Oxford University Press), 1998, pp. 574-662.

[GAR00] Gärdenfors, P., *Conceptual spaces. The geometry of thought.* Cambridge, MA (MIT Press), 2000.

[GAR02] Gardner, J. and Z. Rendon, *XSLT and XPATH. A guide to XML transformations.* Upper Saddle River (Prentice Hall), 2002.

[GES04] G. Geser and J. Pereira (editors), *The future digital heritage space. An expedition report* [online] (DigiCULT Thematic Issue 7), 2004 [cited 19 November 2004]. PDF format. Available from World Wide Web: <http://www.digicult.info/downloads/dc_thematic_issue7.pdf>.

[GIL00] Gilliland-Swetland, A., *Enduring paradigm, new opportunities: the value of the archival perspective in the digital environment.* Washington D.C. (Council on library and information resources), 2000.

[GIL02] Gilchrist, A., 'Thesauri, taxonomies and ontologies – an etymological note'. In: *Journal of Documentation*, vol. 59 (2002), pp. 7-18.

[GON92] González, P., 'Computerisation project for the Archivo General de Indias'. In: P. Doorn, C. Kluts and E. Leenarts (eds.), *Data, computers and the past. VGI Cahier 5.* Hilversum (Verloren), 1992, pp. 52-67.

[GON99] González, P., *Computerization of the Archivo General de Indias. Strategies and results.* Amsterdam (European Commission on Preservation and Access), 1999.

[GON04] González, R. and K. van der Meer, 'Standard metadata applied to software retrieval'. In: *Journal of Information Science*, vol. 30 (2004), pp. 300-309.

[GRE89] Greenough, S. and others, *On the art of fixing a shadow. One hundred and fifty years of photography*. Chicago (National Gallery of Art), 1989.

[GRE02] Greenstein, D. and S. Thorin, *The Digital Library: A biography.* [online] Washington D.C. (CLIR Reports), 2002 [cited 9 November 2004]. PDF format. Available from World Wide Web: <http://www.clir.org/pubs/reports/pub109/pub109.pdf>.

[GRE03] Greenstein, D. and A. Smith, 'Digital preservation in the United States. Survey of current research, practice, and common understandings'. March 2002'. In: *New model scholarship: How will it survive?* Washington D.C. (CLIR Reports), 2003, pp. 46-55. Also available in PDF version from World Wide Web: <http://www.clir.org/pubs/reports/pub114/pub114.pdf> [cited 5 October 2004].

[GRO00] Grout, C.P. *et al.*, *Creating digital resources for the visual arts: Standards and good practice*. Oxford (Oxbow books), 2000.

[GUI00A] *Guides to quality in visual resource imaging*. [online] (DLF, RLG, CLIR) 2000 [cited 24 April 2003]. HTML format. Available from World Wide Web: <http://www.rlg.org/visguides>.

[GUI00B] *Guide to best practice: Dublin Core.* [online] (CIMI: Consortium for the computer interchange of museum information) 2000 [cited 8 June 2004]. PDF format. Available from World Wide Web: <http://staff.dstc.edu.au/andrewg/standards/cimidc/cimidc.pdf>.

[HED95] Hedstrom, M., 'Preserving the intellectual record: a view from the archives. In: Dempsey, L., D. Law and I. Mowat (editors) *Networking and the future of libraries 2: managing the intellectual record.* London (Library Association Publishing), 1995, pp. 179-191.

[HED02] Hedstrom, M., 'The digital preservation research agenda'. In: *The state of digital preservation: An international perspective.* Washington D.C. (CLIR Reports), 2002, pp. 32-37 [cited 22 May 2004]. PDF format. Available from World Wide Web: <http://www.clir.org/pubs/reports/pub107/pub107.pdf>.

[HEE00] Heery, R. and M. Patel, 'Application profiles: mixing and matching metadata schemas' [online]. In: *Ariadne*, vol. 5 (2000) [cited 5 April 2003]. HTML format. Available from World Wide Web: <http://www.ariadne.ac.uk/issue25/app-profiles/>.

[HEE02] Heery, R. and Wagner, H., 'A metadata registry for the semantic web' [online]. In: *D-Lib Magazine*, vol. 8 (2002) [cited 3 June 2003]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/may02/wagner/05wagner.html>.

[HEE03] Heery, R. *et al.*, 'Metadata schema registries in the partially Semantic Web: the CORES experience' [online]. In: *Proceedings of the 2003 Dublin Core Conference*, Seattle, 2003 [cited 5 August 2004]. PDF format. Available from World Wide Web: <http://www.siderean.com/dc2003/102_Paper29.pdf>.

[HIT02] Hitchcock, M.S., *Perspectives in Electronic Publishing: Experiments with a New Electronic Journal Model* [online]. PhD-thesis (Southampton University), 2002 [cited 23 May 2004]. PDF format. Available from World Wide Web: <http://www.ecs.soton.ac.uk/~sh94r/Jnls-research/thesis/thesis-text.pdf />.

[HJE01] Hjelm, J., *Creating the semantic web with RDF.* New York (Wiley), 2001.

[HOD04] Hodge, G. and E. Frangakis, *Digital preservation and permanent access to scientific information: The state of the practice A Report sponsored by the International Council for Scientific and Technical Information (ICSTI) and CENDI US Federal Information managers group.* [online], 2004 [cited 19 September 2004]. PDF format. Available from World Wide Web: <http://cendi.dtic.mil/publications/04-3dig_preserv.pdf>.

[HOL89] Holland, J., K. Holyoak, R. Nisbett and P. Thagard, *Induction. Processes of inference, learning, and discovery*. Cambridge (MIT Press), 1989.

[HOM04] Hommes, B., *The Evaluation of business process modelling techniques*. PhD thesis (Delft University of Technology), Delft, 2004.

[HOG94] Hogenboom, J. and J. van der Voort (editors), *Voor de zoeker: Handleiding voor het registreren en uitwisselen van gegevens over fotocollecties.* Den Haag (NBLC Uitgeverij), 1994.

[HOR99] Horik, R. van, 'ISAD(G) wat kun je ermee? Het digitale stadsarchief van Antwerpen'. In: *Archievenblad*, vol. 103, nr. 9 (1999), pp. 16-17.

[HOR01A] Horik, R. van, 'Archives and photographs: the 'European Visual Archive project (EVA)'' [online]. In: *Cultivate Interactive*, issue 3, (2001) [cited 12 May 2003]. HTML format. Available from World Wide Web: <http://www.cultivate-int.org/issue3/eva/>.

[HOR01B] Horik, R. van, 'Digitalisering van fotomateriaal' In: *Handboek informatiewetenschap voor bibliotheek en archief.* Alphen aan den Rijn (Kluwer / Bohn Stafleu Van Loghum), pp. IV E 300-1-IV E 300-22.

[HOR02] Horik, R. van and K. van der Meer, 'Building blocks for durable metadata of visual sources'. In: P. Isaias (editor) *Proceedings of the 2nd international workshop on new developments in digital libraries NDDL 2002*. Setubal (ICEIS press), 2002, pp. 80-92.

[HOR03] Horik, R. van, 'About the longevity of digital surrogates of historical photographs: the importance of application profiles'. In: Anderson, J., A. Dunning and M. Fraser (editors), *Digital resources for the Humanities 2001-2002. An edited selection of papers.* London (Office for the Humanities computing), 2003, pp. 271-279.

[HOR04] Horik, R. van, H. Koppelaar, K. van der Meer and P. Doorn, 'Permanent pixels: Building blocks for the longevity of digital surrogates of historical photographs'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 128-135.

[HYV02] Hyvönen, E., A. Styrman and S. Saarela, 'Ontology-based image retrieval'. In: Hyvönen, E. and M. Klemettinen (editors), *Towards the semantic web and web services – Proceedings of the XML Finland 2002 Conference.* Helsinki (HIIT), 2002, pp. 15-27.

[IFL98] *Functional Requirements for Bibliographic Records* / IFLA Study Group on the Functional Requirements for Bibliographic Records [online]. Munich (K.G. Saur) 1998 [cited 2 March 2003]. PDF format. Available from World Wide Web: <http://www.ifla.org/ VII/s13/frbr/frbr.pdf>.

[ISO02] *Draft ISO/TC 46 Business plan as evaluated and accepted by the ISO TMB BP Taskforce*, 2002. International Organisation for Standardisation.

[JAC00] Jacobson, R., G. Attridge, S. Ray and N. Axford, *The manual of photography. Photographic and digital imaging*. Oxford (Focal Press), 2000.

[JAN04] Janosky, J. and R. Witthus, 'Using JPEG2000 for enhanced preservation and web access of digital archives - A case study'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 145-149.

[JON96] Jonker, M., M. Boom, P. Reinders and A. van Veen (eds), *Assessing Photographs. Criteria for the assessment of photographic collections*. Rotterdam (Netherlands Photographic Society), 1996.

[JON01] Jones, M. and N. Beagrie, *Preservation management of digital materials.* London (The British Library), 2001.

[JON02] Johnston, D.L., 'A simplified standard method of digital image tonal capture for archival projects'. In: *Proceedings of the IS&T Image processsing, Image Quality, Image Capture (PICS) Conference 2002 PICS Conference*, Portland, Oregon (Society for imaging science and technology), 2002, pp. 210-213.

[JUN04] Jung, K. and Th. Zellmann, 'JPEG2000/Part 6 for scanned documents in archiving applications'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 281-285.

[KAS00] Kashyap, V. and A. Sheth, *Information brokering across heterogeneous digital data: a metadata-based approach*. Boston (Kluwer), 2000.

[KEN96] Kenney, A. and S. Chapman, *Digital imaging for libraries and archives*. Ithaca, NY (Cornell University Library), 1996.

[KEN99] Kenney, A. and L.H. Sharpe, *Illustrated book study: digital conversion requirements of printed illustrations* [online]. Report to the Library of Congress preservation directorate. Contractors: Cornell University Library and Picture Elements inc. (1999) [cited 4 January 2003]. PDF and HTML format. Available from World Wide Web: <http://www.library.cornell.edu/preservation /ill_bk_cover.htm>.

[KEN00] Kenney, A. and O. Rieger, *Moving theory into practice. Digital imaging for libraries and archives*. Mountain View (Research Libraries Group), 2000.

[KIR98] Kirsch, R., 'SEAC and the start of image processing at the National Bureau of Standards'. In: *Annals of the history of computing, IEEE,* vol. 20 (1998), pp. 7-13.

[KLI00] Klijn, E. and Y. de Lusenet, *In the picture. Preservation and digitisation of European photographic collections.* Amsterdam (European Commission on Preservation and Access), 2000.

[KLI03] Klijn, E. (ed.), *SEPIADES. Recommendations for cataloguing photographic collections*. Amsterdam (European Commission on Preservation and Access), 2003.

[KON87] Koningsveld, H., *Het verschijnsel wetenschap. Een inleiding in de wetenschapsfilosofie*. Amsterdam (Boom), 1987.

[KUN01] Kunze, J., 'A metadata kernel for electronic permanence' [online]. In: *Journal of digital information*, issue 2 (2001) [cited 23 August 2004]. PDF format. Available from World Wide Web: <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Kunze/kunze-final.pdf>.

[LAP04] Laplante, P., *Software engineering for image processing systems*, Boca Raton, FL (CRC Press), 2004.

[LAW00] Lawrence, G. *et al.*, *Risk management of digital information: A file format investigation.* [online] Washington D.C. (CLIR Reports), 2000 [cited 14 March 2004]. PDF format. Available from World Wide Web: <http://www.clir.org/pubs/reports/pub93/ pub93.pdf>.

[LAZ01] Lazinger, S. S., *Digital preservation and metadata. History, Theory, Practice*. Englewood, Colorado (Libraries Unlimited), 2001.

[LEF02] LeFurgy, W.G., 'Levels of service for digital repositories' [online]. In: *D-Lib Magazine*, vol. 8 (2002) [cited 13 May 2003]. HTML format. Available from World Wide Web: <http://www.dlib.org/ dlib/may02/lefurgy/05lefurgy.html>.

[LOE01] Loebich, C. and D. Wueller, 'Three years of practical experience in using ISO standards for testing digital cameras'. In: *Proceedings of IS&T Image processing, Image Quality, Image Capture (PICS) Conference*, Montreal, Quebec (Society for imaging science and technology), 2001, pp. 257-261.

[LOR01] Lorie, R., 'Long-term archiving of digital information'. In: *Proceedings of the first ACM/IEEE-CS joint conference on digital libraries, Roanoke, Virginia*, New York (ACM Press), 2001, pp. 346-352.

[LOR02] Lorie, R., *The UVC: a method for preserving digital documents. Proof of concept.* Den Haag (Koninklijke Bibliotheek), 2002.

[LUS02] Lusenet, Y. de, 'Old photographs, new images. The SEPIA program on preservation of photographic collections'. In: *La conservation à l'ère du numérique. Actes des quatrièmes journées internationales d'Etudes de l'ARSAG* Paris (Association pour la recherche scientifique sur les arts graphiques), 2002.

[LYN01] Lynch, C., 'Metadata harvesting and the Open Archives Initiative'. In: *ARL Bimonthly Report* Issue 217 (August 2001), pp. 1-9. Also available in HTML format from World Wide Web: <http://www.arl.org/newsltr/217/mhp.html> [cited 5 October 2004].

[LYN03] Lynch, C. 'Institutional repositories: Essential infrastructure for scholarship in the digital age' [online]. In: *ARL Bimonthly Report*, Issue 226 (February 2003) [cited 11 November 2004]. HTML format. Available from World Wide Web: <http://www.arl.org /newsltr/226/ir.html>.

[MAE02] Maedche, A. *Ontology learning for the semantic web.* Boston (Kluwer Academic Publishers), 2002.

[MAL01] Maly, K., M. Zubair and X. Liu, 'Kepler – An OAI data/service provider for the individual' [online]. In: *D-Lib Magazine*, vol. 7 (2001) [cited 9 September 2003]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/april01/maly /04maly.html>.

[MAT02] Mattison, D., 'Images of history on the Web' [online]. In: *Searcher* issue 5 (2002) [cited 7 February 2003]. HTML format. Available from World Wide Web: <http://www.infotoday.com/searcher /may02/mattison.htm>.

[MCK99] McKemmish, S., G. Acland, N. Ward and B. Reed, 'Describing records in context in the continuum: the Australian recordkeeping metadata schema'. In: *Archivaria*, no. 48 (1999), pp. 3-43. Also available in HTML version from World Wide Web: <http://www.sims.monash.edu.au/research/rcrg/publications/archiv01.htm/> [cited 29 October 2004].

[MEE02] Meer, K. van der, *Documentaire informatiesystemen.* Den Haag (Biblion), 2002 (4th revised edition).

[MIL90] Miller, F., *Arranging and describing archives and manuscripts* Chicago (The Society of American Archivists), 1990.

[MIL00] Miller, P., 'Interoperability What is it and Why should I want it?' [online] In: *Ariadne*, vol. 5 (2000) [cited 8 January 2004]. HTML format. Available from World Wide Web: <http://www.ariadne.ac.uk /issue24/interoperability/>.

[MOO02] Moore, R., 'The San Diego Project: Persistent Objects' [online]. In: *Proceedings of the Workshop on XML as a Preservation Language,* (Urbino, Italy) 2002 [cited 19 May 2003]. Word format. Available from World Wide Web: <http://www.sdsc.edu/NARA /Publications/persistent-objects.doc>.

[MUH02] Mühlberger, G., 'Automated digitisation of printed material for everyone: The metadata engine project' [online]. In: *RLG Diginews*, vol. 6 (2002) [cited 22 November 2004]. HTML format. Available from World Wide Web: <http://www.rlg.org/legacy /preserv/diginews/diginews6-3.html#feature1>.

[MUR94] Murray, J.D. and W. vanRyper, *Encyclopedia of graphics file formats*. Sebastopol, CA (O'Reilly & Associates), 1994.

[MUR04] Murray, R.J., 'JPEG2000 in practice: The effect of image content and imaging system characteristics'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 266-274.

[MRP04] Murphy, E.P., *A review of standards defining testing procedures for characterizing the color and spatial quality of digital cameras used to image cultural heritage.* [online] Rochester, NY (Rochester Institute of Technology, Munsell Color Science Laboratory), 2004 [cited 12 August 2004]. PDF format. Available from World Wide Web: <http://www.cis.rit.edu/museumSurvey/documents /StandardsReview_tp.PDF>.

[OST98] Ostrow, S.E., *Digitizing Historical Pictorial Collections for the Internet.* Washington / Amsterdam (Council on Library and Information Resources), 1998. Also in HTML format. Available from World Wide Web: <http:/www.clir.org/pubs/reports/ostrow /pub71.html> [cited 5 October 2004].

[PAY98] Payette, S., 'Persistent identifiers on the digital terrain' [online]. In: RLG Diginews, vol. 2(1998) [cited 15 October 2004]. HTML format. Available from World Wide Web: <http://www.rlg.org/legacy /preserv/diginews/diginews22.html#Identifiers>.

[PER04] *Persistent identification: A key component of an E-government infrastructure* [online] (CENDI Persistent identification task group) 2004 [cited 11 November 2004]. PDF format. Available from World Wide Web: <http://cendi.dtic.mil/publications/04-2persist_id.pdf>.

[PRE01] *Preservation metadata for digital objects: A review of the state of the art* [online]. A white paper by the OCLC/RLG working group on preservation metadata. January 31, 2001 [cited 23 June 2004]. PDF format. Available from World Wide Web: <http://www.oclc.org/research/pmwg/presmeta_wp.pdf>.

[PRO03] Prom, C., 'Reengineering archival access through the OAI protocols'. In: *Library High Tech*, vol. 21 (2003), pp. 199-209.

[PRO04] *Producer-Archive Interface Methodology Abstract Standard*, [online] published by Consultative Committee for Space Data Systems, CCSDS 651.0-B-1, Blue Book, May 2004 [cited 26 October 2004]. PDF format available form World Wide Web: <http://www.ccsds.org/CCSDS/documents/651x0b1.pdf>.

[QUE04] Quenault, H., 'VERS: Building a digital record heritage'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 2-7.

[RIE04] Rieger, O., 'Implementing a digital imaging and archiving program: technology meets reality'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 191-194.

[REI86] Reilly, J.M., *Care and identification of 19th century photographic prints*. Rochester, NY (Eastman Kodak), 1986.

[REI00] Reich, V. and D. Rosenthal, 'LOCKSS (Lots Of Copies Keep Stuff Safe)'. In: *The new review of academic librarianship,* vol. 6 (2000), pp. 155-161.

[ROB93] Robinson, P., *The digitization of primary textual sources.* Oxford (Office for humanities communication publications), 1993.

[ROS00] Ross, S., *Changing trains at Wigan: Digital preservation and the future of scholarship*. London (National Preservation Office), 2000. Also available in HTML version from World Wide Web: <http://eprints.erpanet.org/archive/00000045/> [cited 5 February 2004].

[ROS03] Ross, S., *Digital Library Development Review*. [online] (National Library of New Zealand), 2003 [cited 18 August 2004]. PDF format. Available from World Wide Web: <http://eprints.erpanet.org/archive/00000050/01/ross_report.pdf>.

[ROT95] Rothenberg, J., 'Ensuring the longevity of digital documents'. In: *Scientific American* (January 1995), pp. 42-47.

[ROT00] Rothenberg, J., *Using emulation to preserve digital documents*. The Hague (Rand-Europe / Koninklijke Bibliotheek), 2000.

[SEA02] Seadle, M. 'METS and the metadata marketplace'. In: *Library Hi Tech*, vol. 20 (2002), pp. 255-257.

[SEA03] Searle, S. and D. Thompson, 'Preservation metadata. Pragmatic first steps at the national library of New Zealand' [online]. In: *D-Lib Magazine*, vol. 9 (2003) [cited 4 February 2004]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/april03 /thompson/04thompson.html>.

[SHA99] Shaw, R., 'A century of image quality'. In: *Proceedings of IS&T Image processing, Image Quality, Image Capture (PICS) Conference*, Savannah, Georgia (Society for imaging science and technology), 1999, pp. 221-224.

[SHA00] Shapiro, S., *Thinking about mathematics. The philosophy of mathematics.* Oxford (Oxford University Press), 2000.

[SHE98] Shepard, T., 'Universal Preservation Format (UPF): Conceptual framework' [online]. In: *RLG Diginews*, vol. 2 (1998) [cited 23 January 2003]. HTML format. Available from World Wide Web: <http://www.rlg.org/preserv/diginews/diginews2-6.html>.

[SIT00] Sitts, M.K. (editor), *Handbook for digital projects: A management tool for preservation and access*. Andover, Massachusetts (Northeast Document Conservation Center), 2000.

[SKY66] Skyrms, B., *Choice and chance. An introduction to inductive logic.* Belmont, CA (Dickenson Publishing Company), 1966.

[SMI04] Smith, A., 'Mapping the preservation landscape' [online]. In: *Access in the future tense.* Washington D.C. (CLIR Reports), 2004, pp. 1-8 [cited 15 December 2004]. PDF format. Available from World Wide Web: <http://www.clir.org/pubs/reports/pub126 /pub126.pdf>.

[SNE02] Snell, J., D. Tidwell and P. Kulchenko, *Programming Web Services with SOAP*. Sebastopol, CA (O'Reilly), 2002.

[SOM01] Sompel, H. Van de and Oren Beit-Arie, 'Open linking in the scholarly information environment using the OpenURL framework' [online]. In: *D-Lib Magazine*, vol. 7 (2001) [cited 20 June 2004]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>.

[SOM04] Sompel, H. Van de, and others, 'Resource harvesting within the OAI-

PMH framework'[online]. In: *D-Lib Magazine*, vol. 10 (2004) [cited 28 December 2004]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/december04/vandesompel /12vandesompel.html>.

[SON79] Sontag, S., *On photography*. London (Penguin), 1979.

[SOW00] Sowa, J. F., *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove (Brooks/Cole), 2000.

[SPI01] Spivak, S.M. and F.C. Brenner, *Standardisation essentials. Principles and Practice*. New York (Dekker), 2001.

[STA04] Stanescu, A., 'Assessing the durability of formats in a digital preservation environment' [online]. In: *D-Lib Magazine*, vol. 10 (2004) [cited 15 November 2004]. HTML format. Available from World Wide Web: <http://www.dlib.org/dlib/november04/stanescu /11stanescu.html>.

[STE02] Steenbakkers, J., 'Preserving electronic publications'. In: *Information services and use,* vol. 22 (2002), pp. 89-96.

[STE03A] Stehno, B., A. Egger and G. Retti, 'METAe – Automated encoding of digitized texts'. In: *Literary and Linguistic Computing*, vol. 18 (2003), pp. 77-88.

[STE03B] Steingrimsson, Ú. and K. Simon, Perceptive quality estimations: JPEG 2000 versus JPEG'. In: *Journal of imaging science and technology,* vol. 47 (2003), pp. 572-585.

[STR97] Strothotte, C. and Th. Strothotte, *Seeing between the pixels. Pictures in interactive systems*. Berlin (Springer), 1997.

[SVE00] Svenonius, E., *The intellectual foundation of information organization*. Cambridge (MIT Press), 2000.

[TAL99] Tally, T., *Avoiding the output blues.* Upper Saddle River (Prentice Hall), 1999.

[TAN98] Tannenbaum, R., *Theoretical foundations of multimedia*. New York (Computer Science Press), 1998.

[THI02] Thibodeau, T., 'Overview of technological approaches to digital preservation and challenges in coming years' [online]. In: *The state of digital preservation: An international perspective.* Washington D.C. (CLIR Reports), 2002, pp. 4-31 [cited 2 February 2003]. PDF format. Available from World Wide Web: <http://www.clir.org/pubs/reports/pub107/pub107.pdf>.

[THO92] Thorvaldsen, G., 'The preservation of computer readable records in the Nordic countries'. In: *History and Computing*, vol. 4 (1992), pp. 201-205.

[UMB98] Umbaugh, S., *Computer vision and image processing. A practical approach using CVIPtools.* Upper Saddle River (Prentice Hall), 1998.

[UND04] *Understanding Metadata* [online], Bethesda MD (NISO Press), 2004 [cited 4 September 2004]. PDF format. Available from World Wide Web: <http://www.niso.org/standards/resources /UnderstandingMetadata.pdf >.

[VEL01] Veltman, K., 'Syntactic and semantic interoperability: new approaches to knowledge and the semantic web'. In: *The new review of information networking*, vol. 7 (2001), pp. 159-183.

[VER04] Verheusen, A. 'Het TIFF-archief van de Koninklijke Bibliotheek'. In: *Historia en Informatica*, vol. 11, nr. 4 (2004), p. 4. Also available in HTML version from World Wide Web: <http://www.historiaeninformatica.org/11_4/04.htm> [cited 19 December2004].

[VRE95] Vreede, G.J. de, *Facilitating organizational change. The participative application of dynamic modelling*. Dissertation Delft University of Technology, 1995.

[WAL03] Wall, G. 'Business model issues in the development of digital cultural content' [online]. In: *First Monday,* vol. 8 (2003) [cited 12 March 2004]. HTML format. Available from World Wide Web: <http://www.firstmonday.dk/issues/issue8_5/wall/index.html>.

[WAI03] Waibel, G., 'Like Russian dolls: nesting standards for digital preservation' [online]. In: *RLG Diginews,* vol. 7 (2003) [cited 16 October 2003]. HTML format. Available from World Wide Web: <http://www.rlg.org/preserv/diginews/diginews7-3.html#feature2>.

[WAI04] Waibel, G. and R. Dale, 'Automatic Exposure: Capturing technical metadata for digital still images'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 260-265.

[WEI96] Weibel, S. and E. Miller, 'Image description on the Internet: Summary of CNI/CLC Image Metadata Workshop' [online]. In: *Annual Review of OCLC Research 1996* [cited 4 June 2003]. HTML format. Available from World Wide Web: <http://www.oclc.org/research/publications/arr/1996/image.htm>.

[WEI02] Weiss, P., *Borders in cyberspace: conflicting public sector information policies and their economic impacts* [online] February 2002 [cited 5 October 2004]. PDF format. Available from World Wide Web: <http://www.weather.gov/sp/Bordersreport2.pdf>.

[WEL02] Wells, L. (editor), *Photography: a critical introduction*. London (Routledge), 2002.

[WIJ04] Wijngaarden, H. van and E. Oltmans, 'Digital preservation and permanent access: The UVC for images'. In: *Proceedings of IS&T 2004 Archiving Conference,* San Antonio, Texas (Society for imaging science and technology), 2004, pp. 254-258.

[WIL98] Williams, D., 'What is MTF … and why should you care?' [online] In: *RLG Diginews,* vol. 2 (1998) [cited 12 February 2003]. HTML format. Available from World Wide Web: <http://www.rlg.org/preserv/diginews/diginews21.html#technical>.

[WIL00] Williams, D., 'Selecting a scanner' [online]. In*: Guides to quality in visual resource imaging.* (DLF, RLG, CLIR) 2000 [cited 3 September 2003]. HTML format. Available from World Wide Web: < http://lyra2.rlg.org/visguides/visguide2.html>.

[WIL02] Wilde, E. and D. Lowe, *Xpath, Xlink, Xpointer, and XML. A practical guide to web hyperlinking and transclusion.* Boston (Addison-Wesley), 2002.

[WIL03] Williams, D., 'Debunking of speckmanship: Progress on ISO/TC42 Standards for digital capture imaging performance'. In: Proceedings *of the IS&T Image processing, Image Quality, Image Capture (PICS) Conference 2002 PICS Conference*, Rochester, New York (Society for imaging science and technology), 2003, pp. 77-81.

[WUE01] Wueller, D. and C. Loebich, 'Practical scanner tests based on OECF and SFR measurements'. In: *Proceedings of IS&T Image processing, Image Quality, Image Capture (PICS) Conference*, Montreal, Quebec (Society for imaging science and technology), 2001, pp. 252-256.

[ZWE92] Zweig, R., 'Virtual records and real history'. In: *History and Computing*, vol. 4 (1992), pp. 174-182.

Normative references

[DCF98] *Design Rule for Camera File System*. Version 1.0 [online]. Japan Electronic Industry Development Association (JEIDA), December 1998 [cited 15 June 2004]. PDF format. Available form World Wide Web: <http://www.exif.org/dcf.PDF>.

[DIG00] *DIG35 Specification. Metadata for digital images*. Version 1.0, August 30, 2000. Digital Imaging Group, Inc.

[DNG04] *Digital Negative (DNG) Specification*. Version 1.0.0.0 [online]. Adobe Systems Incorporated, September 2004 [cited 8 October 2004]. PDF format. Available from World Wide Web: <http://www.adobe.com/products/dng/pdfs /dng_spec.pdf>.

[EXI02] *Exchangeable image file format for digital still cameras. (EXIF)*. Version 2.2 [online]. Japan Electronics and Information Technology Industries Association (JEITA), April 2002 [cited 15 June 2004]. PDF format. Available from World Wide Web: < http://www.exif.org/Exif2-2.PDF>.

[ICC04] ICC.1:2004-04 *Image technology management – Architecture, profile format, and data structure [Revision of ICC.1:2003-09]*. International Color Consortium, April 2004 [cited 5 November 2004]. PDF format. Available from World Wide Web: <http://www.color.org /ICC1V42.pdf>.

[IEC61966-2-2:2003] IEC 61966-2-2: 2003 *Multimedia systems and equipment – Colour measurement and management – Part 2-2: colour management – Extended RGB colour space – scRBG*, International Electrotechnical Commission.

[ISO/TC46:2002] ISO/TC 46: 2002 *Draft ISO/TC 46 Business plan as evaluated and accepted by the ISO TMB BP Taskforce*, International Organisation for Standardisation.

[ISO690-2:1997] ISO 690-2: 1997 *Information and documentation – Bibliographic references – Part 2: Electronic documents or parts thereof*, International Organisation for Standardisation.

[ISO704:2000] ISO 704: 2000 *Terminology work – Principles and methods*, International Organisation for Standardisation.

[ISO3664:2000] ISO 3664: 2000 *Viewing conditions – Graphic technology and photography,* International Organisation for Standardisation.

[ISO8632:1994] ISO/IEC 8632: 1994 *Information technology- Computer graphics metafile. Metafile for the storage and transfer of picture description information. Part 1: Functional specification*, International Organisation for Standardisation.

[ISO10646:1993] ISO/IEC 10646: 1993 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and basic multilingual plane (plus amendments AM1 through AM 7*, International Organisation for Standardisation.

[ISO10918-1:1994] ISO/IEC 10918-1: 1994 *Information technology – Digital compression and coding of continuous-tone still images – Part 1: Requirements and guidelines*, International Organisation for Standardisation.

[ISO10918-2:1995] ISO/IEC 10918-2: 1995 *Information technology – Digital compression and coding of continuous-tone still images – Part 2: Compliance testing*, International Organisation for Standardisation.

[ISO11179-1:1999] ISO/IEC 11179-1: 1999 *Information technology – Specification and standardisation of data elements – Part 1: Framework for the specification and standardisation of data elements*, International Organisation for Standardisation.

[ISO11179-1:2004] ISO/IEC 11179-1: 2004 *Information technology – Metadata registries (MDR) – Part 1: Framework*, International Organisation for Standardisation.

[ISO11179-2:2000] ISO/IEC 11179-2: 2000 *Information technology - Specification and standardisation of data elements – Part 2: Classification of data elements*, International Organisation for Standardisation.

[ISO11179-3:1994] ISO/IEC 11179-3: 1994 *Information technology – Specification and standardisation of data elements – Part 3*, International Organisation for Standardisation.

[ISO11179-3:2003] ISO/IEC 11179-3: 2003 *Information technology – Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes*, International Organisation for Standardisation.

[ISO11179-4:1995] ISO/IEC 11179-4: 1995 *Information technology - Specification and standardisation of data elements – Part 3: Rules and guidelines for the formulation of data elements,* International Organisation for Standardisation.

[ISO11179-4:2004] ISO/IEC 11179-4: 2004 *Information technology – Metadata registries (MDR) – Part 4: Formulation of data definitions*, International Organisation for Standardisation.

[ISO11179-5:1995] ISO/IEC 11179-5: 1995 *Information technology - Specification and standardisation of data elements – Part 5: Naming and identification principles for data elements,* International Organisation for Standardisation.

[ISO11179-6:1997] ISO/IEC 11179-6: 1997 *Information technology - Specification and standardisation of data elements – Part 6: Registration of data elements,* International Organisation for Standardisation.

[ISO12233:2000] ISO 12233: 2000 *Photography – Electronic still picture cameras – Resolution measurements*, International Organisation for Standardisation.

[ISO12234-1:2001] ISO 12234-1: 2001 *Electronic still-picture imaging – removable memory – Part 1: Basic removable-memory module*, International Organisation for Standardisation.

[ISO12234-2:2001] ISO 12234-2: 2001 *Electronic still-picture imaging – removable memory – Part 2: TIFF/EP image data format,* International Organisation for Standardisation.

[ISO14524:2000] ISO 14524: 2000 *Photography – Electronic still-picture cameras – Methods for measuring opto-electronic conversion functions (OECFs),* International Organisation for Standardisation.

[ISO14721:2003] ISO 14721: 2003 *Space data and information transfer systems - Open Archival Information System- Reference Model,* International Organisation for Standardisation. Also published as: *Reference model for an Open Archival Information System (OAIS)*, [online] (Consultative Committee for Space Data Systems), CCSDS 650.0-B-1, Blue Book, January 2002 [cited 5 October 2004]. PDF format. Available from World Wide Web: <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents /pdf/CCSDS-650.0-B-1.pdf>.

[ISO15444-1:2004] ISO/IEC 15444-1: 2004 *Information technology – JPEG2000 image coding system – Part 1: Core coding system*, International Organisation for Standardisation.

[ISO15452:2000] ISO/IEC TR 15452: 2000 *Information technology – Specification of data value domains. Technical Report.* International Organisation for Standardisation.

[ISO15529:1999] ISO 15529: 1999 *Optics and optical instruments - Optical transfer function - Principles of measurement of modulation transfer function (MTF) of sampled imaging systems,* International Organisation for Standardisation.

[ISO15739:2003] ISO 15739: 2003 *Photography - Electronic still-picture imaging - Noise measurements*, International Organisation for Standardisation.

[ISO15836:2003] ISO 15836: 2003 *Information and documentation – The Dublin Core metadata element set*, International Organisation for Standardisation.

[ISO15948:2004] ISO/IEC 15948: 2004 *Information technology – Computer graphics and image processing – Portable Network Graphics (PNG): Functional Specification*, International Organisation for Standardisation.

[ISO16067-1:2003] ISO 16067-1: 2003 *Photography - Spatial resolution measurements of electronic scanners for photographic images- Part 1: Scanners for reflective media,* International Organisation for Standardisation.

[ISO16067-2:2004] ISO 16067-2: 2004 *Photography - Spatial resolution measurements of electronic scanners for photographic images- Part 2: Film scanners,* International Organisation for Standardisation.

[ISO18921:2002] 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.

[ISO19501:2002] ISO/IEC 19501-1: 2002 *Information technology – Unified Modeling Language (UML) – Part 1: Specification*, International Organisation for Standardisation.

[ISO20943-1:2003] ISO/IEC TR 20943-1: 2003 *Information technology – Procedures for achieving metadata registry (MDR) content consistency – Part 1: Data elements*, International Organisation for Standardisation.

[ISO20943-3:2004] ISO/IEC TR 20943-3: 2004 *Information technology – Procedures for achieving metadata registry (MDR) content consistency – Part 3: Value domains,* International Organisation for Standardisation.

[ISO21550:2003] ISO 21550: 2003 *Photography – Electronic scanners for photographic images – Dynamic range measurements*, International Organisation for Standardisation.

[ISO22028-1:2004] ISO 22028-1: 2004 *Photography and graphic technology – Extended colour encodings for digital image storage, manipulation and interchange – Part 1: Architecture and requirements*, International Organisation for Standardisation.

[ISO23081-1:2004] ISO 23081-1: 2004 *Information and documentation – Records management processes – Metadata for records – Part 1: Principles*, International Organisation for Standardisation.

[NISOZ39.87:2002] NISO Z39.87: 2002 *Data dictionary – Technical metadata for digital still images* [online] NISO AIIM. Status: Draft standard for trial use [cited 25 August 2004]. PDF format. Available from World Wide Web: <http://www.niso.org/ standards/resources/Z39_87_trial_use.pdf>.

[TIF92] *TIFF – Revision 6.0 Final June 3, 1992* [online] [cited 2 January 2004]. PDF format. Available from World Wide Web: <http://partners.adobe.com/public/developer/en/ tiff/TIFF6.pdf>.

[XML04] *Extensible Markup Language (XML)* 1.0 Third edition. W3C Recommendation [online] 4 February 2004 [cited 5 October 2004]. HTML format. Available from World Wide Web: <http://www.w3.org/TR/2004/REC-xml-20040204/>.

About the author

René van Horik (born in Aarle-Rixtel, May 1963) started studying Economic and Social History at Nijmegen University in 1982. He graduated in 1987 with a thesis analysing indicators for nineteenth-century proto-industrialisation in the Dutch southern province of Noord Brabant. Once accustomed to the personal computer, which had been introduced by that time, he found it not only to be a convenient word-processor for producing the thesis, but also a potentially useful instrument for storing and processing historical data. Thus, an interest in using ICT in the Historical discipline was aroused.

In early 1989 the author became involved in a pilot project carried out at Leiden University to set up the Netherlands Historical Data Archive (NHDA). From the outset it was clear that the NHDA would serve not only as an archive for digital scientific output in the traditional sense, but also as an expert centre in the field of the application of ICT in the Humanities. In the following years the author carried out research and projects in the area of the conversion of a wide range of types of analogue historical sources into digital versions. In January 1995 the NHDA became an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the author became head of the department 'Digitisation and Consultancy'. In the period 1991–1999 he was involved as a trainer and supervisor in the annual post-doctoral course 'Historical Information Processing', a joint initiative of the NHDA and Leiden University. During his career the author has written a considerable number of articles and taken part in conferences, workshops and other meetings.

The concept of the present PhD dissertation gradually developed after September 1997, when the NHDA became part of the Netherlands Institute for Scientific Information Services (NIWI-KNAW), where the author held a position as senior specialist – digital data archives. Several projects and research activities carried out at NIWI-KNAW by the author addressed the digitisation of, and access to, historical photographic collections. The often labour-intensive and expensive conversion projects require attention to access and usage of digital objects in the long term, but fixed, generally accepted digital preservation solutions are not yet available.

Since July 2005, the author has worked as a theme manager at Data Archiving and Networked Services (DANS), a new national organisation formed to create a research data infrastructure in the Humanities and the Social and Behavioural Sciences. DANS is a joint initiative of the Netherlands Organisation for Scientific Research (NWO) and the Royal Netherlands Academy of Arts and Sciences.

The BETADE research programme of Delft University of Technology provided the author with the opportunity to work on this dissertation in his field of interest on a part-time basis in the period 2000-2004.