

IBM  
long

KB  
term

evaluation

study

010010

authenticity  
in a digital  
environment

010







010010 authenticity 010  
in a digital  
environment

Dr. Raymond J. van Diessen  
and  
Drs. Titia van der Werf-Davelaar

01001011

01000010

Design: Steven L. Stijger  
Published by: IBM Netherlands, Amsterdam  
IBM / KB Long-Term Preservation Study  
Report Series Editor: Dr. Raymond J. van Diessen

Available from:	
IBM External Communications	Koninklijke Bibliotheek
PO Box 9999	PO Box 90407
1000 CE Amsterdam	2509 LK The Hague
The Netherlands	The Netherlands

Title: **Authenticity in a Digital Environment**

ISBN: 90-6259-155-8  
Authors: Dr. Raymond J. van Diessen and Drs. Titia van der Werf - Davelaar  
Date: December 2002  
Copyright: IBM / Koninklijke Bibliotheek

*This study was commissioned by the Koninklijke Bibliotheek,  
National Library of the Netherlands*

# IBM / KB

# long-term preservation study

The National Library of the Netherlands (Koninklijke Bibliotheek, KB) is faced with the problem of preserving large amounts of digital documents for the long term. These documents come from two sources: from media published directly in digital form and from digitizing paper documents. In 2000, the KB and IBM started building an electronic deposit system ("Digital Information Archiving System or DIAS"), the technical core of the infrastructure for KB's e-Deposit for the Netherlands.

From the beginning it was clear that this project could not rely on out-of-the-box solutions alone because up to that time no solution readily addressed both the aspects of large volume and durable storage as well as the long-term preservation requirements. So an IBM / KB Long-Term Preservation Study (LTP Study) was initiated as part of the overall project of developing an electronic deposit system.

The primary objective of the LTP Study was to investigate the functionality required for the long-term preservation (hundreds of years) of the digital information stored in DIAS. This study has resulted in 6 reports: one overview report and five specific reports, each one addressing an important aspect of long-term preservation in its own right.

Participants in the LTP Study:

#### **IBM**

Raymond J. van Diessen  
Raymond Lorie  
Sidney Huiskamp  
Hans Verhoeven

#### **Koninklijke Bibliotheek**

Johan F. Steenbakkens  
Titia van der Werf-Davelaar  
Patricia Alkhoven  
Adriaan Lemmen

#### **RAND Corporation**

Jeff Rothenberg

#### **British Library**

Deborah Woodyard

I would like to thank all the participants for their input and enthusiasm. The results make an important contribution to the development and implementation of dedicated functionality for the long-term preservation of digital information and for guaranteeing long-term access.

Report Series Editor,  
Raymond J. van Diessen

# Titles of the Report Series

## **Number 1: The Long-Term Preservation Study of the DNEP Project - an Overview of the Results**

This report explains the reasons and objectives behind defining the LTP Study as part of the overall project to implement an electronic deposit system. It also provides a quick and general overview of all the study results, which are then elaborated on in the other published reports.

## **Number 2: Authenticity in a Digital Environment**

Authenticity acquires a new meaning in a digital context. Normally objects are physical and their physical characteristics are the main source for defining authenticity. Moreover, authenticity is not a single concept, but involves different aspects that can be associated with an object:

- ∄ A traceable path from the object's origin to its current ownership.
- ∄ Measures and techniques for safeguarding against and/or recognizing modifications.
- ∄ Techniques for establishing the use of original materials.

The problem of digital objects is that in fact they are just conceptual objects. A digital object is a conceptual object to be interpreted (rendered) by executing the digital object in a specific IT infrastructure (hardware & software). This report focuses on defining a framework in which we can define what is actually meant when one speaks of an authentic digital object.

## **Number 3: Preservation Requirements in a Deposit System**

The initial DIAS release only provides basic functionality for preserving and rendering the stored digital objects for the long term. One of the primary responsibilities of the LTP Study is to define the functional requirements of the Preservation Subsystem, which is scheduled for development later. This report identifies requirements of the DIAS Preservation Subsystem so as to provide the services and functions for monitoring the technical environment associated with the digital objects stored in DIAS.

The Preservation Subsystem can be summarized by the following three objectives:

- ∄ Identifying digital objects that are in danger of becoming inaccessible because of changes in technology.
- ∄ Implementing the activities associated with technical preservation.
- ∄ Supplying the requisite technical metadata in order to generate / validate the environments needed during digital object delivery.

## **Number 4: The UVC: a Method for Preserving Digital Documents - Proof of Concept**

Within IBM Research in Almaden, Raymond Lorie was already working on a combined emulation / migration approach to preserve a certain class of digital objects with an approach called the Universal Virtual Computer (UVC).

The main idea consists of archiving a program P along with the data file that decodes the data and returns the information to a future client based on a logical view. The logical view of the data is simple and self-contained enough to be interpreted without any specific software or hardware. Program P is written for the Universal Virtual Computer (UVC) that is general, yet basic enough to continue to be relevant in the future. Given the simplicity of the UVC, it will be relatively easy to write an emulator of the UVC in the future on a real machine of that time. The emulated machine will run the program P and return all data in an easy to understand logical view of the data.

The LTP Study conducted a proof of concept with the KB to test the UVC approach in a library environment. The PDF format was selected because it is the primary data format for electronic publications to be stored in DIAS.

## **Number 5: Managing Media Migration in a Deposit System**

Storage technology obsolescence makes media migration a necessity. Data has to be copied from one storage medium to another on a regular basis. However, the fact that storage technology becomes obsolete is not the only trigger for rewriting previously stored digital objects. All storage media degrade over time and have to be rewritten either on the same medium (refreshing) or on another medium (migration).

Ordinarily media refreshment / migration would be a straightforward process. However, the large amounts of storage associated with an electronic deposit system introduce certain volume-specific requirements. Most electronic deposit systems define their storage capacity needs in several TeraBytes ( $10^{12}$  Bytes). Take a deposit system with 100 TeraBytes of information stored on tape, for example. Let's assume that you want to migrate all this information to an optical storage medium. Current optical storage media have a capacity of around 5 GigaBytes and a write speed of around 4 MegaBytes/second. A quick calculation shows that a complete migration to optical storage would take at least 290 days (100 TeraBytes / 4 MegaBytes per second)!

This report describes the actions to be taken to manage media migration / refreshment effectively within an electronic deposit system, focussing specifically on the media migration issues within DIAS. Potential additional capacity required for media migration might be created by redundancy and parallelism.

## **Number 6: Archiving Web Publications**

More and more Web publications are becoming a primary source of information and will thus be stored as digital objects in DIAS. Web publications have specific characteristics and requirements that DIAS must meet if they are to be archived successfully.

This report investigates the issues and requirements introduced by archiving Web publications and their potential impact on DIAS.



# 01 contents



1/	Summary	1
2/	Authenticity	3
	2.1 Authenticity is Context Dependent	4
	2.2 Authenticity in a Digital Environment	5
3/	Information	7
	3.1 Processing Information	7
	3.2 Information Systems	8
4/	Interpretation	11
	4.1 Designation	12
	4.2 Mathematical Interpretation	12
	4.3 Phenomenological Interpretation	13
	4.4 Total Process	14
5/	Digital Object Interpretation	15
	5.1 Digital Object Interpretation Schemata	15
	5.2 Authenticity Framework	16
	5.3 Maximum Effectiveness	18
	5.4 Authenticity Definition Process	19
6/	Authenticity within an Electronic Deposit	21
	6.1 Deposit Library Preservation Requirements	21
	6.2 The Nature of Electronic Publications	21
	6.3 Deposit Authenticity Principles	22
	6.4 Digital Object Types Encountered within DIAS	25
7/	Conclusions	29
	Appendix A: References	33
	Appendix B: Glossary	35





# summary



Authenticity acquires a new meaning in a digital context. Normally objects are physical and their physical characteristics are the determining factors for defining authenticity. Moreover, authenticity is not a single concept but involves different aspects that can be associated with an object:

- € A traceable path from the object's origin to its current ownership.
- € Measures and techniques for safeguarding against and/or recognizing modifications.
- € Techniques for establishing the use of original materials.

This report focuses on designing a framework by which we can define what is actually meant by an authentic digital object, i.e. what are the criteria used to measure authenticity.

A digital object is a conceptual object to be interpreted (rendered) by executing the digital objects in a specific IT infrastructure (hardware & software). The proposed authenticity framework identifies five global types of interpretation schemata used to render a digital object:

1. Binary Interpretation Schemata.
2. Content Interpretation Schemata.
3. Content Metadata Interpretation Schemata.
4. Structure Interpretation Schemata.
5. Functional Interpretation Schemata.

The framework proposed defines the aspects of a digital object in relation to its interpretation process. For each of these interpretation schemata the organization maintaining the electronic deposit has to determine whether or not it sees the specific interpretation schema as information that has to be preserved. The steps to be taken are:

1. Identify the basic binary schemata used by the IT infrastructure.
2. Identify the different content schemata available for each digital object type.
3. Identify the metadata content schemata, e.g. font type, bold, underline.
4. Identify the structure schemata for each digital object type.
5. Identify the functional schemata supported and their impact on the digital object type.
6. Select the interpretation schemata that must be preserved - in addition to the binary and content interpretation schemata - and evaluated in the context of the objectives and capabilities of the deposit.

The six-step authenticity definition process provides a structured process for discussing and defining the authenticity criteria for each digital object type. It enables one to assess preservation strategies and determine how they affect the authenticity of digital objects. This is an initial attempt to develop and deploy ubiquitous preservation functions in digital environments.



# 2/ 01 authenticity 01000010

Authenticity acquires a new meaning in a digital context. Normally objects are physical and their physical characteristics are the main source for defining authenticity. Authenticity is also not a single concept, but involves different aspects that can be associated with an object:

- € A traceable path from the object's origin to the current ownership.
- € Measures and techniques for safeguarding against and/or recognizing modifications.
- € Techniques for establishing the use of original materials.

Usage and context define how these aspects are defined for individual classes of objects.

*Digital information technology creates significant risks that electronic records may be altered, either inadvertently or intentionally. Therefore, in the case of records maintained in electronic systems, the presumption of authenticity must be supported by evidence that a record is what it purports to be and has not been modified or corrupted in essential respects.*

Assessing the integrity and authenticity of records is crucial to archiving practice. A digital environment exacerbates this requirement. "Digital information technology creates significant risks that electronic records may be altered, either inadvertently or intentionally. Therefore, in the case of records maintained in electronic systems, the presumption of authenticity must be supported by evidence that a record is what it purports to be and has not been modified or corrupted in essential respects." [InterPARES 2001]. In archiving practice, authenticity is documented throughout a document's entire life cycle starting the moment it is created based on the provenance principle. However, this aspect of authenticity in the context of original producer, e.g. the publisher, is outside the scope of this report. The primary goal of this study is to establish a workable framework for defining the authenticity characteristics of digital objects when they are entered in the deposit system. This should be done in a structured manner independent of the digital object types so that a workable preservation process is defined within the deposit system.

## 2.1 Authenticity is Context Dependent

The Night Watch in the Dutch Rijksmuseum, serves as a good example for discussing how authenticity is related to context. Actually it has another title: the 'Company of Frans Banning Cocq and Willem van Ruytenburch'. The picture, painted in 1642, is a group portrait of a division of the civic guard. Rembrandt depicted the group of militiamen in an original way. He did not paint them in neat row or sitting at their annual banquet, rather, he recorded a moment: a group of militiamen have just moved into action and are about to march off.

This painting is a perfect example for illustrating what we mean by authenticity and for showing that the concept is not as clear-cut as one might expect. First of all, one must establish that the painting was really painted by Rembrandt by checking aspects such as the signature on the painting, the age and composition of the colors used, etc. We know from historical records for whom it was painted and we can more or less trace its ownership history up to its current location in the Rijksmuseum.



Figure 2.1 / Night Watch by Rembrandt, 1642

Unfortunately a few years ago the painting was damaged by a mentally disturbed man with a knife and had to be repaired. No one would argue that it is not an original, even though some parts have had to be repaired. We could even say modified. Probably the painting has also been cleaned a few times to bring out the color. None of these actions affect the painting's authenticity, although they all could be classified as modifications.

But is it the same when a building is being renovated or enlarged? Is the building still authentic as once envisioned by the original architect? A building can be regarded as no longer authentic if materials not intended by the original architect are substituted for the original materials. In architecture the use of the applied materials is regarded as an important aspect of the whole building as well as its form.

Within the music industry we believe standards have shifted over time. With current digital recording and filtering techniques it has become possible to "clean-up" an original recording in many ways. A solo of a singer is cleaned up and in the process changes some of the singer's vocal characteristics. However, we still regard it as the original recording! For obvious reasons some of these singers do not like to perform live!

These examples show that what is considered authentic often depends on the type of object under investigation and its suggested usage. There are no uniform rules for deciding what is authentic and what is not.

## 2.2 Authenticity in a Digital Environment

The problem of digital objects is that they are actually just conceptual objects. A digital object is a conceptual object to be interpreted (rendered) by execution in a specific IT infrastructure. The exact characteristics of the actual rendered digital object are thus intimately related to and dependent on the rendering process itself and on the IT infrastructure (hardware & software) used. These are precisely the components that will change most over time. We have to ask ourselves what parts of the digital objects and the supporting rendering environment have to be preserved.

The IBM / KB LTP Study [Diessen and Steenbakkens 2002] has categorized authenticity aspects under the title intellectual preservation, which addresses the integrity and authenticity of the information as originally recorded. It is important to have a workable concept of authenticity in an electronic deposit system. Without it, there would be no way to measure the success of the long term preservation activities.

The following Chapters introduce a framework for defining aspects of a digital object in relation to its interpretation process. The framework will define the authenticity aspects of a digital object in relation to its interpretation process. We will specify a number of different interpretation schema types that will be the basis of our authenticity framework. For each of these interpretation schemata the organization maintaining the electronic deposit has to decide whether or not it sees that aspect as part of the information that has to be preserved.

Authenticity in the context of original ownership, e.g. the publisher, falls outside the scope of this report. Our primary concern is to establish a workable framework for defining the authenticity of digital objects in a structured manner independent of their type, so as to define a workable preservation process.

Deposit libraries like the KB, whose task is to guarantee last-resort access to all published material produced in their own country, are challenged by the possibilities of

today's IT technology. In order to evaluate the effectiveness of the DIAS Preservation Subsystem [Diessen 2002] a workable definition of authenticity must be specified for the digital objects stored in the system.

*The problem of digital objects is that they are actually just conceptual objects. A digital object is a conceptual object to be interpreted (rendered) by execution in a specific IT infrastructure. The exact characteristics of the actual rendered digital object are thus intimately related to and dependent on the rendering process itself and on the IT infrastructure (hardware & software) used.*

# 3/ information 01000010

Information is rapidly becoming one of the most valuable resources in modern society. Throughout history, information exchange has always been an important aspect in terms of progress; however, with the aid of computer technology information exchange has become more effective and widespread today than ever before.

Information technology touches almost every aspect of daily life. Because information technology influences the way we live and function, it changes society itself. Yet there is no foreseeable end to the progress of information technology as it continues its march towards new frontiers. Information technology (IT) enables people to use information in a more advanced manner because it provides the means for enhancing both the processing and the accessibility of the information.

## 3.1 Processing Information

Contrary to most definitions of information, this report will define it as a process. Information is not a concrete substance or concept, but a process. Information materializes as a person or group transforms data into information by interpreting the data in a particular context at a particular time. Thus the correct term for information is the information process [Diessen 1997].

The process of interpretation is initiated by phenomena or symbols, i.e. data. Some data, such as the use of the 365-day calendar, are globally interpreted in the same manner and can thus be perceived as accepted facts. Usually, however, the interpretation process consists of 'attributing meaning' and thus depends on the context as perceived by the person executing the process.

People thus interpret data - i.e. phenomena and symbol systems - differently depending on their background. An analyst will probably interpret the symbol string "model" as a reference to a data model. However, a photographer might think of it as a reference to an individual he/she is working with at that moment. Yet another person may interpret model to be a reference to one of his miniature airplanes. All interpretations are equally valid in their own specific context.

It may seem as if various interpretations of one symbol string (label) as multiple concepts are the result of the ambiguity of the symbol systems, especially when that symbol system is natural language. However, physical phenomena appear to present the same challenges in terms of multiple interpretation. Individuals can interpret a phenomenon such as a sunset on an island differently as well. A tourist just starting his/her vacation might think what a beautiful sunset. In contrast, a tourist on his/her last day of vacation might think, "it is a pity that I have to go home tomorrow".

*Information is not a concrete substance or concept, but a process. Information materializes as a person or group transforms data into information by interpreting the data in a particular context at a particular time.*

## 3.2 Information Systems

Information systems are used to process data. Four basic types of operations are involved in this processing: creation, retrieval, manipulation, and presentation. When creating data, the knowledge collected has to be represented by symbols, which will later be used as input for the information process. In essence, the creation process defines the use of information and, as such, establishes possible presentation schemata. Examples of this are the methods by which so called 'objective' empirical studies can be manipulated to support pre-set, favored conclusions. The manner and the order in which data are presented to a person will influence this person's perception of the data. The information process can be influenced using similar methods.

Data manipulation operations like  $a + b$  will be interpreted as an addition in a calculation context where the  $a$  and  $b$  variables represent numbers. However, in the whole-part context of a mechanical system  $a$  and  $b$  could represent basic electronic components and could represent  $+$  the recursive assembly operator which defines the construction of an electronic system through its basic electronic components. Both interpretations of the operation  $+$  are valid in their individual contexts.

The preceding discussion demonstrates that data manipulation can never be independent of the context in which it is presented. The meaning is something the user attributes to the information process. As a result, information systems can never be neutral data manipulation programs. The information system guides the user towards a preferred mode of interpretation and a specific mode of utilizing the product and its data. Retrieval operations are an important aspect of information manipulation since they are required to retrieve data and related knowledge. Data retrieval is one of the most difficult aspects of managing data. In order for the retrieval process to be successful, it has to be in tune with the contextual 'knowledge' and the user's information needs. Samuel Johnson wrote:

"Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information upon it."

In today's information society, the ability to find information is a highly valuable skill. Even though a great deal of data is available, the difficulty is in selecting the right data in the right context at the right time.

More and more digital objects submitted to deposit libraries are becoming complete information systems in their own right, e.g. a CD-ROM. This makes it harder for the

deposits to separate the software (manipulation, retrieval and presentation functions) from the actual information. This makes it necessary to archive or emulate the complete IT infrastructure needed to use these types of objects as well. Depending on the objectives of the deposit this could be viewed as wasteful and restrictive. The information is only accessible through predefined software, though technology changes will probably provide new ways of access in the future. The deposit library has to decide how important the original user access functionality and the raw information is. How much of the original manipulation and retrieval software is actually needed to effectively interpret the information?



# 4/

# 01 interpretation 000010

In this Chapter we explore the use of conceptual models in the context of information processing. Ever since man gained conscious awareness he has tried to understand himself and the physical world he lives in. The sheer quantity of objects and phenomena in this physical world made it impossible for a single human being to accomplish this task. As time passed, it became impossible to look at each object or phenomenon as an autonomous unit. The overwhelming volume of objects and phenomena required a shift of perspective which entailed a novel approach consisting explaining the behavior of a large class of objects / systems by defining the global knowledge and rules that apply to all members of that class. The limitations of human cognitive powers and the complexity of the world made it necessary to categorize knowledge. We began to look at more efficient ways of exchanging our experiences (information) with one another. Science is one of our most successful endeavors to categorize and codify our experiences.

Traditionally, the field of metaphysics has always struggled with the dichotomy that exists between the substance of an object in the physical world and the manner in which it is perceived by human beings [Agassi 1990]. This dichotomy is a fundamental issue in both modern 'western' philosophy and modern science. The essential question is whether a stable 'objective' substance exists which functions as the source of all appearances an object can manifest. Because of the manner in which people perceive and discuss the physical world, it is commonly believed that mankind cannot identify such an objective substance.

Each individual interprets information, e.g. a book, a digital object or a report. This means that a sonnet can have different effects on different individuals. However, this interpretation process can be broken down into a number of types of interpretation schemata, each of which builds on the knowledge provided by lower level interpretation schemata.

We will first have a look at how the interpretation process has been used in the sciences. The next Chapter will use the same techniques to establish the interpretation framework in the context of digital objects.

Bunge defines four kinds of interpretative relationships; for the purpose of this report these have been consolidated into three relationships: designation, mathematical, and phenomenological [Bunge 1974a, 1974b]. The difference between factual and empirical items is irrelevant in the context of this report; so, for the purposes of this report we have lumped them together under phenomenological interpretation.

## 4.1 Designation

Every formally defined theory can be viewed as a system of symbols. Any symbol system may contain one or more axioms, which are the symbols and strings (combinations of symbols) being supplied without any form of proof. In addition, symbol systems contain rules of production, which regulate the production of new strings. In mathematical symbol systems, these produced strings would be theorems, which represent statements in natural language that have been proven to be correct. In global symbol systems, the definition of theorems is not very strict. This leads to a situation in which the theorems merely state that the string can be produced within the symbol system.

Every conceptual model that has been developed based on a formal theory can be viewed as the production of one or more strings in an associated symbol system. The symbol system, and the conceptual models that can be built utilizing this system, are always partially dependent on the interpretation based on the symbol system's supplied axioms and rules of production, i.e. designation.

One example of designation is the MIU system, which is a symbol system described by Hofstadter [Hofstadter 1987]. The MIU system is a symbol system that only contains the symbols *M*, *I*, and *U*. The generation of strings is defined by one axiom and four rules of inference:

Axioms:

A1) MI is a string of MIU system

Rules of Inference:

variables  $x, y$

R1)  $xI \rightarrow xIU$

R2)  $Mx \rightarrow Mxx$

R3)  $xIIIy \rightarrow xUy$

R4)  $xUUUy \rightarrow xUy$

The MIU system is purely a conceptual symbol system. No references are made to the semantics of the MIU system. It is merely a 'simple' symbol manipulation system. While knowledge exists regarding the manipulation of the strings, their possible meaning remains unclear.

## 4.2 Mathematical Interpretation

Conceptual systems that are solely interpreted by designation only offer entertainment value. This practice can be compared to manipulating Chinese symbols without knowing their meaning. But a Chinese person would attribute additional meaning to these symbols. The same conceptual system can thus be interpreted differently by various groups of people.

The total interpretation (information) process can be viewed as a hierarchy of processes, in which each lower level process gradually becomes a more context-specific interpretation process. Designation is the most global 'abstract' interpretation of a symbol system. Virtually no meaning is attributed to the symbol system and its resulting strings of symbols (conceptual models).

The next step in the interpretation process is mathematical interpretation. Historically, in mathematics there has been a gradual shift away from specialized theories to the global structure of such systems. Algebra serves as one of the earliest examples of this tendency. Algebra is essentially concerned with performing 'algebraic' operations on elements of a set. Gradually, the importance of the set of actual mathematical entities diminished and the emphasis shifted to the form of the operations themselves. These associations between entities in the set are referred to as laws of composition.

The simplest algebraic structure is a magma. Let  $E$  be a set. A mapping  $f$  of  $E \times E$  into  $E$  is called a law of composition  $E$ . Under this law the value  $f(x,y)$  of  $f$  for an ordered pair  $(x,y) \in E \times E$  is called the composition of  $x$  and  $y$ . Often the law of composition ( $f$ ) is represented in infix forms such as:  $x \dot{\cup} y$  or  $x \cup y$ .

Another well-known algebraic structure is a group, which entails some additional characterizations compared to the law of composition. A group is a magma with a compositional law ( $\dot{\cup}$ ) which assigns to each ordered pair  $x \dot{\cup} y$  ( $x,y \in E$ ) another element of  $E$ , which complies to the following rules with  $x,y,z$  in  $E$ :

- (i) associative  $(x \dot{\cup} y) \dot{\cup} z = x \dot{\cup} (y \dot{\cup} z)$
- (ii) left identity element  $e$  for which all  $x \in E$   $e \dot{\cup} x = x$
- (iii) left invertible for each  $x \in E$  there is an element  $y \in E$  with  $y \dot{\cup} x = e$

A mathematical interpretation is defined as the 'mathematical' specialization of a more abstract formal system. The abstract mathematical system group can be interpreted in many different specific mathematical systems: the real numbers when performing multiplication, the complex numbers when performing subtraction, or a  $n$ -dimensional vector space when applying addition. All of these mathematical 'specialized' theories are groups.

Yet, a mathematical interpretation continues to place the system in a conceptual environment. Without the use of additional semantics, there is still no interpretation to link the conceptual models produced by the formal system to the physical universe. The importance of abstract mathematical systems is the identification of global underlying system structures. Linking specific systems to global mathematical systems will demonstrate the similarities and basic characteristics of these specialized systems. They define recurring patterns.

## 4.3 Phenomenological Interpretation

A phenomenological interpretation is the final step in a process which bridges the gap between conceptual models, based on some formally defined symbol system, and the physical universe. A phenomenological interpretation associates the faceless entities within the conceptual model to objects and phenomena in the physical universe. Conceptual models would be useless without this final step in the total interpretation process.

We will illustrate a phenomenological interpretation with an example.

Theory:	Horse Race
Primitives:	$S, \cup$
Axiom1:	$x,y,z \in S$ $(x \cup y) \cup z = x \cup (y \cup z)$
Axiom2:	$S$ is collection of horses
Axiom3:	$\cup$ means joining a race, i.e. $x \cup y = x$ joins $y$ in a horse race

The first axiom defines  $\langle S, \cup \rangle$  mathematically as a semigroup. The other two axioms define the phenomenological interpretation, i.e. the semantics assigned to the symbols and operators of the semigroup. The first semantic assumption identifies the referents in the theory as horses. The second stipulates that  $\cup$  represents participating in a race.

Phenomenological interpretations are always present even though the related axioms are often not explicitly stated. Making the phenomenological interpretation explicit facilitates a better understanding of the conceptual models constructed within the theory.

## 4.4 Total Process

Within sciences the total interpretation process is divided into three separate interpretations. These interpretations follow the process from an abstract theory through a specific mathematical model to a target system. Both the abstract theory and the specific mathematical theory can be represented by symbol systems which define the conceptual models (symbol strings) to be produced. The model of a theory is the set of all symbol strings that can be produced by the theory. A conceptual model based on a theory is a subset of all potential models that can be defined by the theory. The mapping is bi-directional and ensures that different conceptual models constructed with each interpretation can be mapped to one another.

Every 'formal' conceptual model is based on a certain formal symbol system. This formal symbol system consists of a set of mathematical theories and designation rules. The final phenomenological interpretations can be based on three different types of formal theories:

1. Abstract theory  $\Downarrow$  Target systems.
2. Specialized theory  $\Downarrow$  Target systems.
3. Part of a theory's model  $\Downarrow$  Target systems.

The fit between the abstract theory and the target system determines the kind of phenomenological interpretation. If a good fit can be directly established between the abstract theory and the target system, a specialized theory will not be necessary. Often all the possible constructs (strings) of a model are not subjected to phenomenological interpretation. Only a specific part of the model is used in reference to the target system.

The next Chapter will transpose the information given above onto the digital object interpretation process. Keep in mind that the final interpretation is always made on the basis of one's individual context. It cannot be forced into a single unified interpretation.

# 5/ 01 digital object 1000010 interpretation

Interpretation schema types similar to those described in Chapter 4 can be identified in the process for interpreting a digital object. The identification of the different interpretation schemata associated with rendering a digital object will make it possible to define authenticity requirements in a structured framework. This framework of different interpretation schemata types can then be examined in relation to the objectives of the electronic deposit.

## 5.1 Digital Object Interpretation Schemata

We will show the decomposition of a digital object type into different interpretation schemata using a simple example: ASCII text file. Figure 5.1 summarizes the complete interpretation process.

The designation process is handled by the hardware and media preservation efforts. The hardware translates physical characteristics (magnetic media and memory characteristics) into conceptual objects called bits. These bits are then structured by another interpretation schema into bytes. Bytes or multiples of a byte are the main components for storing data in current Von Neumann-based computer systems, although specialized signal processors also allow other schemata down to the level of a single bit. We will refer to these categories of interpretation schemata as binary interpretation schemata.

The binary values of the bytes in the case of an ASCII text file are interpreted to be characters of the alphabet according to a fixed predefined interpretation schema, i.e. the ASCII standard. A single interpretation schema for the content of the digital object does not always have to exist within one digital object type. An MS Word file - the format in which this report was originally produced - contains at least two content interpretation schemata, one for text and one for images. These types of schemata will be referred to as content interpretation schemata.

The final two interpretation schema types are interpretation schemata performed by the individual himself. The first is the mental model mapping to identify and group characters into English language concepts and sentences. The last step is a phenomenological interpretation relating the concept to the person's experience base, in this example a boat.

Even this simple example shows the potential for variances in interpretation. The authors love sailing so we immediately associated the English language concept "BOAT" with the "mental" picture of a sailboat. Without this picture others might have thought of a cruise ship, a tanker, a steamship or an aircraft carrier. Even making the definition within the English language interpretation schema more precise by specifying sailboat would still leave plenty of room for variety in the final phenomenological interpretation, e.g. a racing sailboat or a traditional sailboat, with either one or two masts.

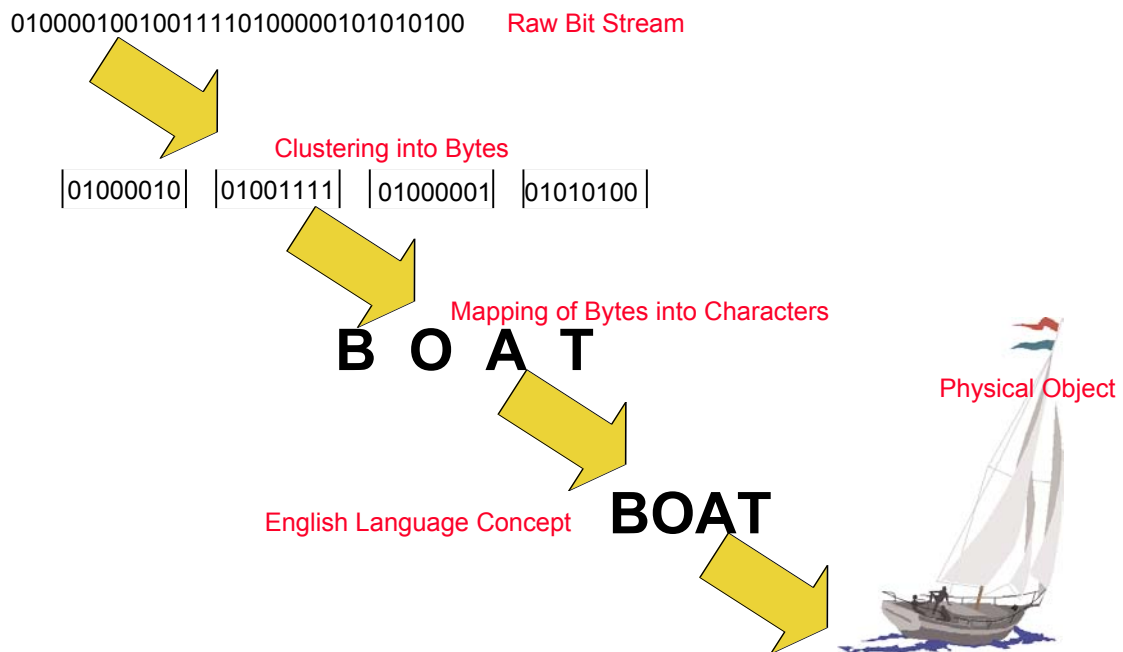


Figure 5.1 / ASCII text interpretation schemata

## 5.2 Authenticity Framework

In addition to the simple interpretation schema types identified in the ASCII example (binary and content interpretation schemata) three other types of interpretation schemata can be identified: content metadata, structure and functional schemata:

### 1. Binary Interpretation Schemata

Define how physical characteristics, most of which are currently magnetic, are translated into bits and further segmented into units of specific bit lengths.

### 2. Content Interpretation Schemata

Define how specific content types are translated into higher level more human-oriented specification concepts, e.g. ASCII, bitmap picture, sound, video, etc.

### 3. Content Metadata Interpretation Schemata

Define additional information / characteristics associated with particular content data

elements. In the case of the ASCII example, codes for bold or underline are regarded as a separate interpretation schema. In this case they say something about the physical format in which certain string of letters must be represented.

#### 4. Structure Interpretation Schemata

Relate different content elements to one another. They provide the means for consolidating different content elements into one coherent aggregate digital object. These structure schemata may include the flexibility to define multiple options within one aggregate. However, the specifications themselves are static and are defined in combination with the content schemata.

#### 5. Functional Interpretation Schemata

This class of schemata includes all application logic used to create, modify, retrieve, delete and render the digital object on a specific IT infrastructure.

Sometimes content metadata and structuring information are not specified explicitly but are enforced by the operation of the associated application software. In these cases it might be necessary to physically render the object in order to extract part of the structuring and content metadata.

Ideally no additional information should be contained in the final transformation of a digital object rendered by some peripheral device. However, often the effort to separate the two has not been made when the digital object was created in its software environment (application). Take the example of the periodic table in which the position of the elements in the tables specifies additional information regarding the characteristics of the specific element or element group, Figure 5.2.

I	II	IIIb	IVb	Vb	VIb	VIIb	VIIIb	Ib	IIb	III	IV	V	VI	VII	0		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H	Li	Be									B	C	N	O	F	Ne	
Na	Mg										Al	Si	P	S	Cl	Ar	
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac**	Rf	Db	Sg	Bh	Hs	Mt	110	111	112	113					
Lanthanides *				Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Actinides **				Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Figure 5.2 / Periodic table defined with non-spatial font

The table above was created with a specific word processor using a non-spatial font type (e.g. Courier) and spaces. In this case, the characteristics of the application and its rendering of information on a peripheral device present relevant information regarding the contents of the digital object. When displayed or printed using a spatial font the information shown in Figure 5.2 would appear something like Figure 5.3.

I	II	IIIb	IVb	Vb	VIb	VIIb	VIIIb	Ib	IIb	III	IV	V	VI	VII	0		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H									He								
Li	Be								B	C	N	O	F	Ne			
Na	Mg								Al	Si	P	S	Cl	Ar			
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac**	Rf	Db	Sg	Bh	Hs	Mt	110	111	112	113					
Lanthanides *			Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	
Actinides **			Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	

Figure 5.3 / Periodic table displayed with spatial font

The absence of the formatting characteristics used to cluster the table into groups makes the presentation of the periodic table of elements in Figure 5.3 incomprehensible. The example shows how important the layout characteristics can be in accurately exchanging information. It also shows that this type of information should be defined explicitly in a separate metadata interpretation schema and not implicitly through the font type used. The periodic table of elements should really have been clustered with the aid of metadata table definitions and tagged accordingly.

Another example is printing an MS Word document using MS Word. This could lead to different results depending on the continent on which the digital object is printed. The US uses the letter paper format, while Europe uses the A4 paper format. The two rendered versions will not appear the same if only the characteristics of the rendered digital object are used. The distribution of the different sections across the individual pages will be different. However, is the rendering process really the essence of the MS Word file or is it just the content and the relations between different content parts that define the real information? In the latter case the rendering of the digital object would be something that could be changed based on individual preference.

## 5.3 Maximum Effectiveness

The effectiveness is maximized when the essence of digital objects to be preserved for the long term is self-contained in the content, the content metadata and the structure schemata to the greatest possible extent. Dependence on functional schemata always directly links the digital object to a specific piece of application logic and/or hardware that has to be maintained over time. This not only limits the digital object's ability to take advantage of functional improvements potential future information technology capabilities may offer; it also makes the authenticity issue dependent on specific functional characteristics and hardware configurations.

The fundamental question is whether the content itself is the main building block or whether this building block is a composite including the functionality to create, modify, delete and render the digital object. We recognize that for some digital objects, like most present day CD-ROMs, no separation can currently be made between the content and the

structure schemata and the functional schema (the application). Therefore, there will always be cases for which some or all of the functional schemata have to be included in the preservation process.

Most publishers of digital objects are not aware of the issues associated with long term preservation. Long term preservation therefore has a low priority. Even worse, some publishers consciously make content dependent up a specific IT infrastructure in order to secure their market share. Most e-book initiatives are set-up this way. But digital objects are fast becoming important items in everyday life. People take their photos and videos in digital format, important letters or e-mails are stored digitally, people keep their financial bookkeeping on their PCs; we could go on and on. At the moment these people are not worried about the accessibility to their valued data because we are still at the start of this technology innovation cycle. However it would be nice to still be able to show 15 years from now the digital photos you took yesterday. This is not a trivial challenge at the moment. More resources will be spent for providing the infrastructure to preserve digital objects for the long term with increased awareness and greater urgency. One of the aspects will be a more balanced approach to the separation of actual content from the functional interpretation schemata (application logic). The logical view of data used by the UVC is a good example of the direction in which this is headed [Lorie 2002].

*The effectiveness is maximized when the essence of digital objects to be preserved for the long term is self-contained in the content, the content metadata and the structure schemata to the greatest possible extent.*

Even inherently dynamic digital object types, like video or sound, can be defined with static structure schemata. Digital video is nothing more than a set of bitmap images arranged in sequence, which have to be displayed evenly in a certain time frame. Sound is a collection of wave samples, which together with the sample rate define the characteristics of the sound wave. Even complex and functionally intensive examples such as relational databases or spreadsheets could be defined statically. However, for some digital object types, like CD-ROMs, it will also be necessary to include part or all of the functional schemata associated with the digital objects.

## 5.4 Authenticity Definition Process

Ultimately every electronic deposit has to define its own authenticity criteria for the different digital object types it contains. At one end of the spectrum a deposit might only

be interested in preserving the core content of the stored digital objects. At the other end of the spectrum a deposit might want to preserve the publisher's complete dissemination environment, with features such as graphic design, branding and advanced search capabilities. Examples of such a publisher's environment are online information delivery services such as Elsevier's Science Direct or Academic Press's IDEAL service.

Every electronic deposit has to perform a number of steps to define its authenticity principles by relating aspects defined in the authenticity framework to the deposit's objectives and capabilities:

1. Identify the basic binary schemata used by the IT infrastructure.
2. Identify the different content schemata available for each digital object type.
3. Identify the metadata content schemata, e.g. font type, bold, underline.
4. Identify the structure schemata for each digital object type.
5. Identify the functional schemata supported and their impact on the digital object type.
6. Select the interpretation schemata that must be preserved - in addition to the binary and content interpretation schemata - and evaluated in the context of the objectives and capabilities of the deposit.

The preservation of the binary and content schemata will always have to be guaranteed by an electronic deposit system in order to make the digital object at least legible. Some structure schemata will also be needed if the digital object type contains more than one content type.

# 6/

# 01 authenticity 01000010 within an electronic deposit

## 6.1 Deposit Library Preservation Requirements

All custodians need to address the issue of authenticity and integrity in the digital environment and they need to do so explicitly. They need to do so because digital objects cannot be accessed over the long term without undergoing some intentional transformation to make them accessible. In other words a digital original does not exist. A digital object is never "authentic". This makes it all the more important to be able to evaluate the authenticity of a digital object: how authentic are the properties of a digital object? How much authenticity has been lost during the transformation of a digital object?

By looking at digital objects generically with the aid of the proposed authenticity framework we can evaluate how authenticity impacts their generic properties. This can help digital heritage custodians make the right choices in light of their own specific objectives and capabilities.

The deposit library represents such a specific application area. After several years of digital deposit practice, preservation research and experiments at the National Library of the Netherlands, insight into the nature of electronic publications and electronic publishing has grown. This section will combine the authenticity framework with the KB's existing insight to position its authenticity criteria within the authenticity framework.

## 6.2 The Nature of Electronic Publications

The KB distinguishes between two types of electronic publications: document-like publications and executable publications.

*By looking at digital objects generically with the aid of the proposed authenticity framework we can evaluate how authenticity impacts their generic properties. This can help digital heritage custodians make the right choices in light of their own specific objectives and capabilities.*

Document-like publications consist of entities that are very similar to those of printed publications. They embody data content, structure and presentation. They require reader functionality for viewing, scrolling and searching for specific words. The readers are part of the application software layer. PDF, HTML, XML and SGML are typical examples of document-like publication formats. But image and sound formats also fall into this category.

It could be argued that reader, viewer and sound-rendering functionality is universal and not specific to any data format. In other words document-like publications are not dependent on their original reader software, because any future reader can provide the required functionality. An interesting aspect of this particular type of digital object is the way in which content, structure, rendering and other support functionality can be distinguished as separate entities. By contrast, a printed book contains all the entities in one: you cannot discard the paper without the content, the pages and the layout. If a digital object can easily be decomposed into the elements that constitute it and some parts are more difficult to preserve than others, the question arises as to which parts should be preserved in the original state and which ones can be replaced by newer technology.

Executable publications are stand-alone digital objects that only require an operating system to be executed. Content, structure, rendering and other support functions are blended into one program that, when activated, manifests itself as one integrated entity. The content parts of such publications, usually appearing in proprietary file formats, are not re-usable within other contexts. Educational software, computer games and Web animations are typical examples of executable publications. It is generally recognized that these publications pose a greater preservation challenge than document-like publications. Emulation is considered one of the most promising technologies for the preservation of executable publications.

## 6.3 Deposit Authenticity Principles

The KB has identified a number of authenticity principles in light of specific KB objectives and capabilities. The authenticity framework discussed in Chapter 5 has helped to facilitate this discussion. Together they have resulted in the definition of the requisite authenticity / preservation support for all the major digital object types encountered in DIAS. Each of these is discussed in more detail in the following sections. Some of the

information in this Chapter has already been published during the Victorian Association for Library Automation (VALA) conference in 2002 [Werf 2002].

### 6.3.1 For Reading and Viewing Only

Deposit copies of electronic publications need only be viewed in a document reader environment as opposed to edited and re-used in a document processing environment. This principle conforms to the way publishers make their publications available. If publishers do provide processing functionality, in exceptional cases the deposit library should also try to make future re-use of data possible. In most cases however, publishers tend to disseminate their products using consumer market standards, such as PDF and HTML. Only in rare instances do they provide dedicated reader software or do they impose the use of a specific reader version. Often publishers can cater to several output formats, which leaves the deposit library with a choice.

The library should provide the reader environment with the appropriate functionality (view, scroll, page down, print, download, word searches). In this sense it seems perfectly legitimate for the deposit library to attempt to provide generic viewing functionality over time and to opt for an UVC-like logical view of data approach [Lorie 2002].

### 6.3.2 Outside the Dissemination Environment

Another principle, related to the previous one, is that the publisher's dissemination environment, with features such as graphic design, branding and advanced search capabilities, is not considered an intrinsic part of the deposited publication. [Steenbakkens 2002] The publisher's own online information delivery services such as Elsevier's Science Direct or Academic Press's IDEAL service thus fall outside the scope of the deposit system. The deposit copy is considered an autonomous published entity that should be definable and identifiable outside its dissemination context as well. This allows deposited publications to be archived in a separate deposit environment with its own search functionality that supports the use of the deposit collection.

The separation of archive versus dissemination environment ensures:

- € Agreement on well-defined and identifiable published entities to be deposited.
- € That the deposit environment does not compete with the added value of the publisher's dissemination environment (indexing, searching, usage rights enforcement mechanisms, etc.).
- € That the deposit system develops its own search environment and preservation functionality and caters to its user's needs over time.

In their proposed approach for a collaborative project funded by the Andrew W. Mellon Foundation, Yale University Library and Elsevier Science have also highlighted the usefulness of the distinction between content and functionality, which enables the separate development of value-added functionality for dissemination purposes and for archival purposes [Yale University & Elsevier Science 2000].

### 6.3.3 Web Publications

Distinguishing between dissemination and preservation environments raises the issue of what to do about Web publications. To what extent can (and should) Web pages be preserved in their dissemination context? How can we define and delimit Web publications for preservation purposes? Are Web sites to be considered publications in their own right or are they publisher dissemination environments? Should a deposit

environment containing Web publications support hyperlinks across Web publications? Should it provide Web search engine functionality? Should it reflect the functionality of the Web as it develops over time?

The initial position taken by the KB on these issues, again as part of the long-term preservation study, is that Web archiving has a different objective from a deposit consisting of electronic publications. Web archiving has grown to mean harvesting and preserving Web pages from the Internet, with the objective of safeguarding the Web and its history for future generations. While Web publications can and should be added to the deposit collection, Web archiving is outside the scope of the deposit library because it preserves much more than just publications: it preserves snapshots of shopping malls and e-bazaars, provides glimpses into the ongoing work of scientific communities and traces of online civic participation. It is an interesting strategy within the broader framework of safeguarding our cultural heritage - but strictly speaking Web archiving is not part of the mission of a deposit library.

*Web archiving has grown to mean harvesting and preserving Web pages from the Internet, with the objective of safeguarding the Web and its history for future generations. While Web publications can and should be added to the deposit collection, Web archiving is outside the scope of the deposit library because it preserves much more than just publications.*

In general, Web publications are document-like publications. Some typical aspects of Web-publications such as hyperlinks, embedded advertisements and interactive buttons with help and feedback functions, require special attention.

Hyperlinks are functional in the Web. One have to maintain their functionality; otherwise a Web page reads like a Table of Contents. Internal links are all intrinsic parts of the publication and should not pose a problem in terms of preservation. The external links are more problematic. They tend to be less informational than in the print world, as URLs are increasingly used in place of full bibliographical citations. This aspect is narrowly related to the whole URI issue on the Web. It has been recognized as an organizational issue, where organizations need to assume their responsibility for generating well-behaved and persistent identifiers that ultimately resolve into locators. Deposit libraries can play an important role in this as providers of last resort locators [Werf 1999].

Advertisements, interactive buttons and other extrinsic features present in a Web publication can be considered a part of this. After all, print publications also carry advertisements. Deposit libraries have never asked publishers to submit copies of journals without advertisements.

## 6.4 Digital Object Types Encountered within DIAS

The Deposit of Netherlands Electronic Publications (DNEP) is a last resort deposit where, ideally at least, the content of all electronically produced information in The Netherlands can be accessed. It does not pretend to be a historical archive showing the functional characteristics of past and present information technology. The technical core of the DNEP is the Digital Information Archiving System (DIAS).

The objective of the KB is to preserve Dutch cultural heritage. Knowledge is codified by symbol systems (digital objects) and interpreted by the information process. We would argue that in most cases it does not matter how the information is rendered. Does Einstein's famous equation  $E=MC^2$  lose any of its meaning when printed in Courier font type as opposed to Times? This is a view shared by most current e-book publishers who often leave font size and type decisions to the individual reader.

Therefore we would like to restrict the addition of the content metadata and structure interpretation schemata to the minimum needed to interpret the digital object. Where possible we try to completely avoid including any functional interpretation schemata at all, although within some of the current digital object types, e.g. certain CD-ROMs, this cannot be avoided.

The following sections detail the authenticity definition process for the most frequently encountered digital object types currently being archived in DIAS. We will not address the binary interpretation schemata in these discussions because reading the actual bit stream is one of the core responsibilities of the DIAS system.

### 6.4.1 PDF

Most of the current publications submitted to KB are in the PDF file format from Acrobat (Adobe) [Adobe 2000].

PDF is defined by four basic interpretation schemes:

#### € **Objects**

A PDF document is a data structure composed of a small set of basic types of data objects.

#### € **File structure**

The PDF file structure determines how objects are stored in a PDF file, how they are accessed, and how they are updated. This structure is independent of the semantics of the objects.

#### € **Document structure**

The PDF document structure specifies how the basic object types are used to represent the components of a PDF document: pages, fonts, annotations, and so forth.

#### € **Content streams**

A PDF content stream contains a sequence of instructions describing the appearance of a page or other graphic entity. These instructions, while also represented as objects, are conceptually distinct from the objects that represent the document structure and are described separately.

We already noted that most PDFs fall into the reading and viewing category. Within this category the content can be viewed as separate from the functional support. One could argue that many of these documents are still printed on paper, even if they are kept in electronic form. For these documents, the way they appear to the human eye is the only consideration. In other words, the e-form is only used for compact storage, but not for additional functionality. For such documents, it would seem reasonable to archive only the image. However, as the number of documents increases, the problem of retrieving the right document(s) becomes harder and harder. Indices on titles or keywords are helpful, but clients, who are getting used to the Internet functionality will very quickly ask for more. It becomes reasonable then to consider that the text itself (in a logical view of the ASCII data) should also be made available to a future application. The same argument can be made for other PDF capabilities such as bookmarks and some of the fields the PDF file contains.

The proof of concept for PDF conducted with the UVC was based on the approach described above [Lorie 2002]. Some issues like the use of specialized content interpretation schemata, e.g. mathematical and chemical formulas, still needs to be addressed. They are generated by mathematical descriptions that more closely resemble a program execution. We have to see whether these should also be maintained in a mathematical format or whether they could be archived as bitmap images. If the mathematical format must be archived for these items, the associated functional interpretation schemata will also have to be preserved. However this could be done with the aid of the UVC emulator to guarantee that the items can be executed on future information technology environments.

*We would argue that in most cases it does not matter how the information is rendered. Does Einstein's famous equation  $E=MC^2$  lose any of its meaning when printed in Courier font type as opposed to Times?*

#### 6.4.2 Bitmap Images

The bitmap images submitted to the DAIS system for preservation are supplied in one of the *de facto* image standards: BMP, GIF, TIFF or JPEG. In addition to these formats the KB also use the Mr. Sid image format that is used to scale image on the Web.

The KB's map collection is one of the major sources of digitized images. In 1808 the KB acquired the majority of the collection of Mr. Joost Romswinkel, consisting of approximately 22-24,000 book volumes and about 9-10,000 maps, plans, etc. This started a collection that has grown over the years to become one of the finest map collections anywhere. Currently the library is working to make the images available electronically as well through a digitizing project.

The primary differences between the individual formats are their ability to compress the data. The KB's focus is thus on separating the content interpretation schema from specific application functionality. The logical view of data provided by the UVC is one mechanism that could be used for this purpose. It will describe the elementary components of an image, i.e. pixels with their associated red, green, blue values, in a technology-independent format. In this way it will provide long-term access to the image without the need for any functional preservation.

Functionality for displaying the images is not very complex and is readily available so current and future applications will provide similar functionality. Potentially the logical view of the data will have to be translated into the *de facto* formats utilized by the applications being used at the time of viewing.

### 6.4.3 Web Publications

In terms of the Internet, the KB focus is on Web publications. Web publications differ from general Web pages in the sense that they resemble normal publications (article, report, books) and also have a publication process associated with them. This excludes discussion pages, theme pages, opinion pages and e-business applications (electronic shopping malls, etc.). Web publications can be divided into two categories: static and dynamic. Static pages are based on HTML pages and do not need any functional support other than the Web server to deliver the pages to the browser. Dynamic Web pages are constructed on the fly by combining information feeds from different information systems and databases.

The HTML pages could be archived using a logical view of HTML. At a minimum this logical view should be able to represent the internal URL references included in the Web publication.

The dynamic pages are more dependent on their functional interpretation schemata (the application supplying the information) and therefore are currently not addressed by the KB. One possibility would be to archive the image of dynamic Web pages although the authors are aware that potentially interesting information embedded in the underlying function schemata would be lost. The advantage of this approach is that the image can be preserved along the lines described for PDF.

### 6.4.4 CD-ROMs

Most CD-ROMs are executable publications that require an operating system to be executed. Content, structure, presentation and functionality are blended into one program that, when activated, manifests itself as one integrated entity. The component parts of such publications, usually in the form of proprietary file formats, are not re-usable within other contexts. Educational software and computer software are typical examples of executable publications. The functional and content interpretation schemata are interwoven and cannot easily be separated.

The complete rendering environment (software and hardware) has to be preserved in this case. Emulation is considered one of the most promising technologies for preserving these executable publications. Although in theory this is feasible, few flexible large-scale emulation projects have been conducted. Some very successful dedicated emulation software has been written: DOS on Window95 or Windows on MacOS. However, to support the long-term preservation objectives of a deposit system, more efficient versions of these emulators must be developed.



# 01 conclusions 01000010

Authenticity is a difficult concept in a digital context. Normally objects are physical and their physical characteristics are the main source for defining authenticity. Authenticity is also not one single concept but involves different aspects that can be associated with an object:

- € The traceable path from the object's origin to its current ownership.
- € Measures and techniques for safeguarding against and/or recognizing modifications.
- € Techniques for establishing the use of original materials.

A digital object is a conceptual object to be interpreted (rendered) by executing the digital objects in a specific IT infrastructure. The proposed authenticity framework identifies five global types of interpretation schemata used to render a digital object:

1. Binary Interpretation Schemata.
2. Content Interpretation Schemata.
3. Content Metadata Interpretation Schemata.
4. Structure Interpretation Schemata.
5. Functional Interpretation Schemata.

The first two types always have to be preserved in order to retrieve the physical bit stream and interpret the basic content. This report has shown how the individual interpretation of the content often also depends on the last type of interpretation schemata: the functional. In this case the application logic used to render the digital object has to be explicitly preserved, which introduces the additional complexity of keeping old software for old hardware available over time. A potential solution for this problem is emulation, but the complexity involved and the lack of actual experience mean that this approach involves considerable risks.

*The six-step authenticity definition process provides a structured process for discussing and defining the authenticity criteria for each digital object type. It makes it possible to evaluate preservation strategies as well as to assess how these strategies affect the authenticity of digital objects.*

If the object can be preserved without any functional schemata, the generally applicable UVC logical view approach supports a technology-independent specification of content, content metadata and structure schemata. The support for the last two schema types depends on the specific object type under investigation and the specific objectives and capabilities of the deposit system.

The six-step authenticity definition process provides a structured process for discussing and defining the authenticity criteria for each digital object type. It makes it possible to evaluate preservation strategies as well as to assess how these strategies affect the authenticity of digital objects. This is an initial step towards developing and deploying ubiquitous preservation functions in digital environments.





# 01 appendix a: 01000010 references

[Adobe 2000]

Adobe Systems Incorporated, *PDF Reference*, Second Edition, Adobe Portable Document Format, version 1.3, Addison-Wesley, 2000.

[Agassi 1990]

Agassi, J., *Ontology and its Discontent*, in *Studies on Mario Bunge's Treatise*, Weingartner, P., and Dorn, G.J.W., eds., pp. 105-122, Rodopi, Amsterdam, The Netherlands, 1990.

[Bunge 1974a]

Bunge, M., *Treatise on Basic Philosophy - Semantics I: Sense and Reference (Volume 1)*, D. Reidel Publishing, Dordrecht, The Netherlands, 1974.

[Bunge 1974b]

Bunge, M., *Treatise on Basic Philosophy - Semantics II: Interpretation and Truth (Volume 2)*, D. Reidel Publishing, Dordrecht, The Netherlands, 1974.

[Diessen 1997]

Diessen, R.J. van, *Model-Driven Object-Oriented Development of Systems: A Behavioural-Oriented Approach*, Hilversum, The Netherlands, 1997.

[Diessen 2002]

Diessen, R.J. van, *Preservation Requirements in a Deposit System*, IBM / KB Long-Term Preservation Study Report Series Number 3, December, 2002.

[Diessen and Steenbakkens 2002]

Diessen, R.J. van and Steenbakkens, J.F., *The Long-Term Preservation Study of the DNEP Project - an Overview of the Results*, IBM / KB Long-Term Preservation Study Report Series Number 1, December, 2002.

[Hofstadter 1987]

Hofstadter, D.R., *Gödel, Escher, Bach: An Eternal Golden Braid*, Penguin Books, Middlesex, 1987.

[InterPARES 2001]

Authenticity Task Force. InterPARES Project, *Requirements for assessing the authenticity of electronic records*, Draft for public comment. July 2001.

[Lorie 2002]

Lorie, R., *The UVC: a Method for Preserving Digital Documents - Proof of Concept*, IBM / KB Long-term Preservation Study Report Series Number 4, December, 2002.

[Steenbakkers 2002]

Steenbakkers, J.F., *Preserving electronic publications*, ICSTI / CODATA / ICSU Seminar on preserving the record of science (in press), IOS Press, Amsterdam, 2002.

[Werf 1999]

Werf, T. van der, *Identification, location and versioning of Web resources*, URI discussion paper, DONOR Report, March 1999.

[Werf 2002]

Werf, Titia van der, *Our digital heritage: how authentic should it be?*, e-volving Information Futures, VALA Conference Proceedings, vol. 1; pp. 285-293, Victorian Association for Library Automation Inc., Melbourne 2002.

[Yale University & Elsevier Science 2000]

*Proposal for a digital preservation collaboration between the Yale University Library and Elsevier Science*, Version 4, 30 September 2000.

# 01 appendix b: 1000010 glossary

**Authentic** means "worthy of acceptance or belief as conforming to or based on fact" and is synonymous with the terms genuine and bona fide.

**Authenticity** is defined as "the quality of being authentic, or entitled to acceptance.

**Bona fide** implies good faith and sincerity of intention.

**Digital Information Archiving System (DIAS)** is the core of the KB's electronic deposit system. Version 1 has been developed by IBM and was released in October 2002.

**DNEP:** In September 2000 the KB and IBM Netherlands signed the final contract which initiated the project "Depot voor Nederlandse Electronische Publicaties" (DNEP) [Deposit for Dutch Electronic Publications] to design and implement DIAS with a Long-Term Preservation Study as an integral part of the total effort.

**Genuine** implies actual character not counterfeited, imitated, or adulterated [and] connotes definite origin from a source.

**Information** often is incorrectly used to refer to data, but information is synonymous with information process, see Information Process.

**Information Process** is the interpretation of observed phenomena and data in a particular context at a particular time by a person or a group.

**KB:** The National Library of the Netherlands (Koninklijke Bibliotheek, KB).

**Logical view of the data:** a view of the data that is easily understandable because it follows the way the user normally thinks about the data, rather than the internal representation often designed for efficiency.

**PDF format:** format designed by Adobe, as part of the Acrobat products, and used to describe the appearance of a page when printed or displayed.

**Program Preservation:** process that ensures that the behavior of a program used today can be re-enacted on a different system in the future.

**Schema:** a formal definition of the logical view of the data.

**Universal Virtual Computer** or UVC is a virtual machine specially designed to develop programs today that will be able to run on a future machine by simply writing an emulator of the UVC in the future.

**XML:** Extensible Markup Language.





IBM  
long

KB  
term

preservation  
study

