WHAT DO WE MEAN BY 'PRESERVING DIGITAL INFORMATION'? TOWARDS SOUND CONCEPTUAL FOUNDATIONS FOR DIGITAL STEWARDSHIP

BY

SIMONE SACCHI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Allen H. Renear, Chair
Associate Professor Jerome P. McDonough
Professor Carole L. Palmer
Dr. Adam Farquhar (British Library)

ABSTRACT

Digital preservation is fundamental to information stewardship in the 21^{st} century. Although much useful work on preservation strategies has been accomplished, we do not yet have an adequate conceptual framework that articulates precisely and formally what preservation actually is. The intention of the account provided here is to bring us closer to this goal. Following an initial analysis of the concept of preservation as it occurs in ordinary discourse around digital stewardship, several influential preservation models are analyzed, identifying both useful insights and problems. A framework of interrelated concepts is then developed that analyzes the challenges of long term digital stewardship through the lens of information communication. Successful digital stewardship is understood as reliable, mediated, intentional communication with an emphasis on the agents involved in the process and the roles they play in supporting the intended flow of information through time and inevitable changes in the underlying mediating communication technology. The complex notion of the digital object, commonly considered the persistent unit of digitally–communicated information, is unpacked into its fundamental abstract and concrete components, avoiding the common category mistakes that pervade digital preservation discourse and impede a clear understanding of the nature of preservation. This conceptual framework makes use of the conceptual machinery of Situation Theory [Devlin, 1995] and the Gricean theory of meaning [Grice, 1957, Grice, 1968]. The notion of an interpretive frame [Dubin et al., 2011] is adopted here to model the contingent mapping between the fundamental components involved in the representation of information and extended with the notion of a constraint

(from Situation Theory) to clarify the role of agent intentionality in the process of establishing the appropriate mappings that ultimately support the successful communication of units of information. This agent–based intentional perspective not only captures the social and contextual nature of successful digital stewardship, but also promises to support a finer grained analysis of preservation expectations from different stakeholders and the potential practical strategies to fulfill them. This research is intended as a contribution to the overall digital preservation agenda by bringing us closer to sound conceptual foundations for the long–term stewardship of our digital scientific and cultural heritage.

ACKNOWLEDGMENTS

The approach taken in this project is based on work initially carried out by David Dubin as part of the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIPP), and then further evolved in the Data Conservancy Data Concepts group (DCDC) at the Center for Informatics Research in Science and Scholarship (CIRSS) at Illinois, funded by the NSF (grant no: 0830976) as part of the Data Conservancy. Principal members of that group during that period were David Dubin, Allen Renear, and Karen Wickett, as well as myself.

This approach is described in the following papers:

Dubin in: Unsworth, J. and Sandore, B. (2010). Echo depository-phase 2: 2008-2010 final report of project activities

Dubin, D., Wickett, K. M., and Sacchi, S. (2011). Content, format, and interpretation. In Usdin, B. T., editor, *Proceedings of Balisage: the Markup Conference 2011*, volume 7 of *Balisage Series on Markup Technologies*, Montreal, Canada

Sacchi, S., Wickett, K. M., Renear, A. H., and Dubin, D. S. (2011b). A framework for applying the concept of significant properties to datasets. In *Proceedings of ASIS&T 2011: the 74th Annual Meeting of the American Society for Information Science and Technology*, volume 48, New Orleans, LA

Sacchi, S., Wickett, K. M., Renear, A. H., and Dubin, D. (2011a). One thing is missing or two things are confused: An analysis of the OAIS Representation Information. In

7th Internation Digital Curation Conference (IDCC), Bristol, UK

Sacchi, S. and Wickett, K. M. (2012). Taking modeling seriously [in digital curation]. In *Research Challenges in Digital Preservation - iPRES 2012*, Toronto, ON, Canada

Wickett, K. M., Sacchi, S., Dubin, D. S., and Renear, A. H. (2012). Identifying content and levels of representation in scientific data. In *Proceedings of ASIS&T 2012: the 75th Annual Meeting of the American Society for Information Science and Technology*, volume 49, Baltimore, MD

I am grateful to them individually as well as collectively for this rich intellectual collaboration and friendship.

Extreme gratitude goes to Allen Renear, my advisor, for his continuous support throughout my Ph.D and for helping me grow as a researcher.

I wish also to thank the other members of my dissertation committee: Jerry McDonough for our many exciting intellectual exchanges, Carole Palmer, former CIRSS director, for making me part of the CIRSS family, and Adam Farquhar, for his availability and precious feedback on my ideas.

My most special thanks go to Claudia, the love of my life, for her limitless patience and invaluable support during every stage of my Ph.D.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The notion of *preserving digital information* is a fundamental concept in recent digital and data stewardship agendas [Task Force on Archiving of Digital Information et al., 1996, Duerr et al., 2004]. Definitions of digital preservation and digital curation directly appeal to this idea. Expressions like "preserving digital information" [Hedstrom, 1997, Waugh et al., 2000, Chen, 2001] or "maintaining digital information" [CCSDS, 2002, Webb, 2003] are also routinely used in the literature and scholarly discourse to refer to the ultimate goal of digital preservation efforts.[1]

Despite the broad adoption of technical terms, the discourse we use to approach the challenges of long-term stewardship —whether in the scientific literature or the hallways of libraries and laboratories— does not reflect a conceptual framework that is sufficiently clear and consistent to be serviceable for resolving these challenges.

The problem has multiple roots, but at its heart is that the notion of *preservation* itself, derived from the traditional library, archive, and museum practice of "minimizing deterioration or damage" of physical artefacts, appears to be a metaphor that misrepresents what is really going on in digital stewardship scenarios, where there is no identifiable thing that is being literally preserved.

Describing stewardship expectations, strategies, and goals in terms of preserving something cannot deliver a precise account of what it means for something 'digital' to be maintained over time, nor properly represent how specific expectations of different communities

[1] See also [Yakel, 2007].

are fulfilled in the practice. Although some approaches to conceptualizing preservation do attempt to transcend *literal preservation*, none of them have been able to fully articulate an alternative account that is sufficiently clear and precise.

Part of the problem is that most theories of preservation make critical use of terms such as 'digital object', 'digital material', and 'digital information' to describe the objects intended to be preserved, but provide these terms with only a loosely defined semantics. In addition, technical terms like 'file' or 'bitstream', which are borrowed from computer science, are not necessarily used in a well–defined technical sense.

In what follows we will begin by demonstrating that literal interpretations of expressions such as 'preserving a digital object' or 'preserving digital information' are not plausible and so they will be deeply inadequate for guiding the development of a formal theory of digital preservation.

Works within the digital preservation community do provide insights that help us move beyond literal interpretation and to understand these expressions more usefully and charitably. However, they do not individually deliver precise and complete accounts and most show critical flaws.

Building on the insights of this prior work, as well as models developed by the Data Conservancy Data Concept (DCDC) research group at Illinois, this dissertation appeals to the Gricean theory of meaning [Grice, 1957, Grice, 1968, Grice, 1969] and to the ontology and formal machinery of Situation Theory [Devlin, 1995, Devlin, 2006] to develop consistent conceptual foundations for digital preservation based on a robust plausible ontology and that promise to be useful for meeting the challenges of digital stewardship.

The basic approach reflects the notion that preservation efforts must "validate the communication from the past" and "enable communication with the future" [Moore, 2008], although it also considers digital preservation itself a form of communication [Mois et al., 2009]. Conditions for successful digital 'preservation' are understood in terms of conditions

to enable a sustained and reliable information communication through time and potential changes in technology.

This project focuses on the preservation of *propositional information*,[2] starting with the meaning assigned to a digital artefact in the intentional act of its 'creation', and accounting for the influence of stewardship expectations, intents, and strategies. In explaining what successfully 'preserving information' really is, we provide a framework for understanding when and why potential communication failure might happen and how to avoid those failures.

Although limiting our treatment to *propositional* information, we of course recognize that there are other features of digital materials that are significant and that must also be accounted for in successful preservation; for example, materials with visual characteristics that cannot be reduced to propositional information (e.g. digital pictures, moving images, digital art, etc.) or material with interactive components (e.g. video games and other interactive media). However we think that a unified conceptual framework that precisely describes how propositional information is represented and communicated in preservation scenarios is a substantial step towards a complete theory of preservation.

The application of such a framework to the digital domain reveals how expectations and goals in digital stewardship can be analyzed —and the means for achieving them— by (a) enabling the use of precise terminology and concepts and (b) revealing why, if information is considered the ultimate target of preservation efforts, we reach different decisions in different contexts.

This framework is also intended to be utilized as a high–level analytical tool in support of preservation activities —e.g. choice of preservation actions, proper metadata characterization— and systems analysis and design in digital preservation.

Ultimately, the goal of this research is to move forward the agenda of developing sound

[2]We take a traditional approach that considers *propositions* to be abstract things which can be the objects of propositional attitudes (such as belief or doubt), the bearers of truth values, and the language–independent entities that are the meanings of those sentences (or other symbol structures) expressing them. For a more extensive treatment, see the entry 'Proposition' in the Stanford Encyclopedia of Philosophy [McGrath, 2012].

conceptual foundations for digital stewardship.

1.1 Research problems

The following research problems are addressed in this research.

RQ1: What do we mean by 'preserving digital information'?

Clearly, we already have some notion of what it is to *preserve digital information*. However, it is difficult to articulate that notion in a clear and precise manner. A successful answer to RQ1 will be a conceptual analysis [Furner, 2004b] of preserving information. This will include both the achieved formal definition and a discussion demonstrating that the definition is correct.

RQ2: What are the problems with current digital preservation models?

There are several conceptual models of digital preservation that are sophisticated, influential, and insightful. What do they get right? What are their flaws?

RQ3: Is there an ontologically sound conceptual framework for digital preservation that incorporates the insights of existing models but avoids their problems?

This research question can be answered by presenting such a framework and then arguing that it is both successful at clearly modeling preservation scenarios and avoiding the problems other models have.

1.2 Organization of chapters

The dissertation is organized according to the following chapters.

Chapter 2

In Chapter 2 we present the methods adopted for this research and relevant conceptual and analytical frameworks this research builds upon.

Chapter 3

In Chapter 3 we show how definitions of *preservation* have evolved to address the evolving nature of information–carrying objects, culminating in definitions of digital preservation where the emphasis is on maintaining access and not preserving objects.

Chapter 4

In Chapter 4 we take a closer look at the broadly adopted notions of *preserving the bits* and *preserving information*, noting how preservation in its literal sense does not apply to the preservation scenarios being described, nor it is required for successful preservation. We then suggest alternative accounts.

Chapter 5

In Chapter 5 we explore three influential approaches to digital preservation that more accurately describe the complex nature of digital material. Despite their many valuable insights, none of these approaches deliver complete and ontologically precise accounts. We also noted that the notion of *preserving access to digital materials* can be more precisely understood from a communication perspective: preserving access to digital materials is about sustaining interpretation and communication processes.

Chapter 6

In Chapter 6 we elaborate on the previous topics by developing a set of interrelated concepts useful to understand how information is represented and communicated. The emerging conceptual framework builds on previous work by the Data Conservancy Data Concepts group at Illinois [Dubin et al., 2009, Dubin et al., 2011, Wickett et al., 2012, Sacchi et al., 2011a,Sacchi et al., 2011a,Sacchi et al., 2011b,Sacchi and McDonough, 2012] and is informed by Paul Grice's theory of meaning [Grice, 1957, Grice, 1968] and Keith Devlin's Situation Theory ontology [Devlin, 1995, Devlin, 2006].

Chapter 7

In Chapter 7 we apply the conceptual framework previously developed as a lens to better understand prototypical issues and opportunities in digital stewardship. Particular emphasis is given to the communication of information sustained by information–carrying artefacts and the mediation roles of the agents involved. We conclude noting that successful digital preservation shall be effectively understood as *successfully sustained reliable communication through time and across changes in the mediating digital technology.*

CHAPTER 2

METHOD AND GROUNDING FRAMEWORKS

2.1 Introduction

This chapter describes the methods we use in our analysis as well as the general conceptual and analytical framework that is applied.

2.2 Method

We take a broadly formal and analytical approach towards understanding the problem of digital preservation, drawing on the principles of *ontology–driven conceptual modeling* [Guarino, 1995,Guarino, 2002,Guarino and Welty, 2002] and the method of *conceptual analysis* [Furner, 2004a] to first clarify core concepts.

Nicola Guarino [Guarino, 1995, Guarino, 2002] uses the phrase "ontology–driven conceptual modeling" to describe the approach adopted here towards developing precise conceptual foundations for digital stewardship. This approach, also called "ontological modeling" [Guizzardi et al., 2003], is characterized by the use of formal languages to specify the conceptualization of a domain and falls within the mainstream of contemporary *formal ontology* as practiced in analytic philosophy [Hofweber, 2013], artificial intelligence [Gruber et al., 1993], and informatics for the natural sciences [Madin et al., 2007, Compton et al., 2012].

The conceptual perspective adopted here will be largely, but not entirely, 'descriptive' (rather than 'revisionary') [Strawson, 1963] with respect to its fundamental orientation, reflecting the "cognitive bias" that is a typical of conceptual work in information science:

> "We do not commit to a strictly referentialist metaphysics related to the intrinsic nature of the world: rather, the categories we introduce here are thought of as cognitive artifacts ultimately depending on human perception, cultural imprints and social conventions (a sort of 'cognitive' metaphysics). We draw inspiration here from Searle's notion of 'deep background' [18], which represents the set of skills, tendencies and habits shared by humans be- cause of their peculiar biological make up, and their evolved ability to interact with their ecological niches [9]. The consequences of this approach are that our categories are at the so-called mesoscopic level, and they do not claim any special robustness against the state of the art in scientific knowledge: they are just descriptive notions [21] that assist in making already formed conceptualizations explicit. They do not provide therefore a prescriptive (or "revisionary" [21, 15]) framework to conceptualize entities. In other words, our categories describe entities in a post-hoc way, reflecting more ore less the surface structure of language and cognition." (Citations from the original text [Gangemi et al., 2002])

In this particular context, we begin by formalizing insights and shared "conceptual schemes" [Gangemi et al., 2002] as they emerge from current models and conceptual approches in digital preservation, and then, where problems arise, attempt to determine the minimal revisions that will solve the described problems.

Although counterintuitive results will be avoided if possible, substantial or even radical proposals will be made if they seem to be necessary to resolve contradictions, avoid unnecessary complications, and improve parsimony.

Analytical methodologies have been developed to support this modeling approach like the influential OntoClean [Guarino and Welty, 2002], which applies notions used for ontological analysis in philosophy. Two sets of principles from OntoClean will be applied throughout our analysis: the principles of *Essence and Rigidity* and the principles of *Identity and Unity*.

2.2.1 Essence and rigidity

Guarino encourages ontology developers to distinguish between types and roles [Guarino and Welty, 2000] and provides a metaproperty, *rigidity*, that functions as a partial criterion for identifying this distinction.

A property is *rigid* if and only if everything that has that property has it necessarily. Every instance of a class necessarily holds it. This notion provides the ground to draw the powerful distinction between *types* and *roles* [Guarino and Welty, 2000]. A traditional example of this distinction is that between *person* and *student*. The class *person* is a type because the property of *being a person* is rigid —a person is a person *necessarily*: if an individual is a person, it has always been a person and cannot cease to be a person. On the other hand, *student* is role a person enters into, because *being a student* is not a rigid property: an individual most likely has not always been a student and eventually will cease to be a student.

The notion of *rigidity* and the distinction between types and roles have proved to provide deep insights into how to shape and make more robust conceptual models and ontologies [Guarino and Welty, 2000].

These criteria can be used to show that the entities identified in a conceptual model are not really independent *types*, but rather actual *roles* that other entities enter into in particular contingent social contexts. [Renear and Dubin, 2007]. The role/type distinction is fundamental to the analysis presented in this dissertation, allowing us to not only to distinguish ontological kinds from role, but surfacing and emphasizing the essential role played by human action and intention in the conceptualization of successful preservation.

2.2.2 Identity and unity

"In general, identity refers to the problem of being able to recognize individual entities in the world as being the same (or different), and unity refers to being able to recognize all the

parts that form an individual entity" [Guarino and Welty, 2002].

Identity conditions are criteria used to determine "circumstances in which something that is apparently seen as one entity is actually two (or more)" in order to avoid *conflation of entities* leading to *category mistakes* in assigning properties to things. An example of analysis indirectly applying this principle is the distinction between the Group 1 entities —Work, Expression, Manifestation, and Item— in the FRBR model [IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998].

Unity conditions are criteria used to determine the mereology of a thing and assess whether part/whole relationships are properly characterized. Understanding compositionality is particularly important to correctly model complex digital resources and the relationships between those components and the information a digital resource is intended to carry. We will show how composition relationships are sometimes abused and improperly applied in the characterization of digital resources.

Within the context of our analysis, the notions of *identity* and *unity* play a dual role:

1. They inform the analysis of current preservation models.

2. They serve as general principles to define the necessary and sufficient conditions for successful digital preservation. Notions like *integrity*, *authenticity* and *persistence* through time —commonly considered critical components of robust preservation efforts— all bear some kind of relationship with the general notions of identity and unity, or to "weaker", but still related, relationships like *equivalence*.

2.3 Grounding conceptual and analytical frameworks

This research is informed and builds upon much previous conceptual work in information science and communication theory.

In particular, it adapts Paul Grice's theory of meaning [Grice, 1957] and Keith Devlin's

work in Situation Theory [Devlin, 1995, Devlin, 2006] and expands the work we did as Data Conservancy Data Concept research group (DCDC).

It also builds on the many insights of previous conceptual approaches to digital preservation developed within the library and archival communities (see Chapter 4 & 5).

2.3.1 The Basic Representation Model (BRM)

The Basic Representation Model (BRM) [Wickett et al., 2012] was based on early work by David Dubin [Sandore and Unsworth, 2010] and further developed as part of the activities of the Data Conservancy Data Concept group (DCDC), a research group led by Allen Renear and David Dubin and based at the Center for Informatics Research in Science and Scholarship at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.[1]

BRM was developed to better understand and model levels of representation in scientific data. However, BRM is general enough to be effectively applied to any situation where we need to analyze how information is represented.

In [Wickett et al., 2012] we defined the three entity types —*Propositional content*, *Symbol Structure*, *Patterned Matter & Energy*— and three relationship types —*Is Expressed By*, *Is Encoded By*, *Is Inscribed In* that together compose BRM.

Propositional content "In our model propositions appear as the language–independent content expressed by symbol structures. In this sense intended propositions may be defined as all and only those things that are either possibly true or possibly false. That is, they are the proper subjects of truth values. The symbol structure that expresses a proposition may also be considered true or false, but only in a derivative sense:

[1]The Data Conservancy Data Concept (DCDC) group was funded by the National Science Foundation as part of the Data Conservancy, a multi–institutional NSF funded project (OCI/ITRDataNet 0830976) hosted at Johns Hopkins University Sheridan Libraries. Carole Palmer, former CIRSS Director, was the co-PI on the project.

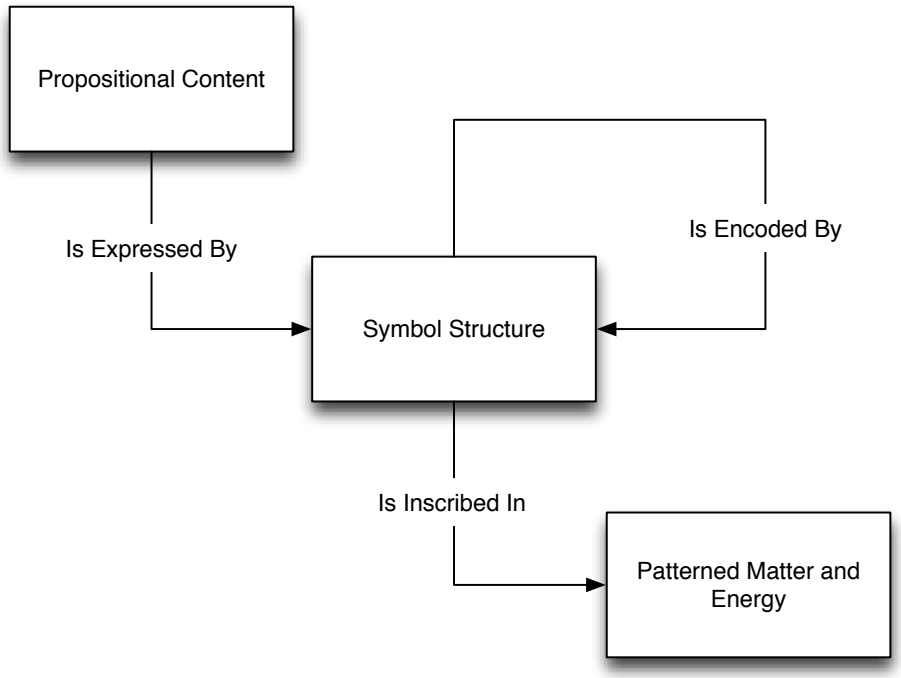

Figure 2.1: Basic Representation Model diagram

derivatively "true" if the proposition it expresses is true, and derivatively "false" if the proposition it expresses is false. A common alternative account of propositions defines them as the proper objects of epistemic attitudes, such as belief or doubt. For our purposes these two accounts of proposition may be considered coextensive: the class of things that can be true or false is identical with the class of things that can be the object of epistemic attitudes. Although the significant role of propositions in our model is as the expressed content of symbol structures, the definitions just given allow propositions to exist independently of symbol structures." [Wickett et al., 2012]

Symbol Structure "In our model symbol structures are abstract arrangements of symbols that, in a given context, express propositions. Individual symbols themselves are the atomic components of symbol structures. Although the symbol structures in our examples are in some language with a determinate semantics, our model allows symbols and symbol structures to express different propositions in different languages or differ-

ent contexts. Examples of abstract objects that can serve as symbol structures include graphs, relations, and sequences, along with more familiar kinds of symbol structures like strings of characters." [Wickett et al., 2012]

Patterned Matter and Energy "Whereas both propositions and symbol structures are abstract objects, patterned matter and energy is a concrete quantity of matter and energy that manifests a physical arrangement that is the physical inscription of an (abstract) symbol structure. In order for a digital object to effectively communicate information, there must be some instantiation of the symbol structures in a physical medium that an agent can interact with." [Wickett et al., 2012]

The relationship types in BRM are defined as follows:

Is Expressed By "Every meaningful digital object will use symbol structures to express propositions. For instance, a digital object may use RDF triples to express propositions about species occurrence. We use the is Expressed By relationship type for this technical sense of 'express'. The Is Expressed By relationship type represents the fact that the propositional content of a digital object is understood as being expressed by a symbol structure that is the primary expression —the Primary Symbol Structure— for that content in a particular context. Is Expressed By represents a general relationship that is instantiated between specific propositional content and a specific symbol structure." [Wickett et al., 2012]

Is Encoded By "A digital object will typically map the symbol structures that express propositions into other symbol structures. We call this mapping from symbol structure to symbol structure an encoding of one symbol structure by (or into) another. For instance, a digital object may map RDF triples into the XML/RDF serialization language. Or those same triples might be encoded in the N3 serialization language. In each case we have the same Primary Symbol Structure —the RDF triples that express

propositional content— but a different encoding of that primary symbol structure. Symbol structures that are encodings of other symbol structures may in turn be encoded by still other symbol structures. For instance the N3 symbol structure may itself be encoded in a UTF-8 byte sequence. Unpacking the encoding levels provides a more complete and consistent way to represent what changes when digital objects undergo transformations, like format migrations." [Wickett et al., 2012]

Is Inscribed In "The *Is Inscribed In* relationship type represents the fact that a particular symbol structure is represented in a physical medium through a mapping between the symbol structure and a particular concrete arrangement of matter and energy." [Wickett et al., 2012]

The important aspect of BRM is its attempt to identify primitive kinds: entities that do comply with the distinction between *types* and *roles* that is essential for proper and consistent conceptual modeling.

Notions like *Propositional Content*, *Symbol Structure*, and *Patterned Matter & Energy* reflect established ontological kinds extensively investigated in philosophical ontology and are adopted as primitives in our discourse about digital preservation. We note that the first two are abstract and the third is concrete; this distinction is discussed further below.

BRM provides the foundational backbone for the *representation stack* of information adopted here.

2.3.2 Gricean theory of meaning

Paul Grice's theory of meaning was first presented in a 1957 Philosophical Review article [Grice, 1957], and later revised [Grice, 1968, Grice, 1969].

In particular, his analysis of *speaker meaning* runs as follows:

(G1) a utters x intending an agent b to form the belief that p

(G2) a intends that b recognizes (1)

(G3) a intends that b forms the belief that p at least partly because b recognizes (1)

In this context we will appeal to Grice's analysis of *speaker meaning* to develop an analytical account of what it means to *intentionally* inform.

2.3.3 Situation Theory

Originally developed by Jon Barwise and John Perry in the 80s, Situation Theory has evolved through the contribution of many scholars, most research being conducted at the Center for the Study of Language and Information (CSLI), an interdisciplinary research center at Stanford University.

In particular, we appeal here to Keith Devlin 'flavor' of Situation Theory as presented in his influential book "Logic and Information" [Devlin, 1995] and in later papers (e.g. [Devlin, 2006]).

According to Devlin, Situation Theory is

> "a set of mathematically-based tools to analyze, in particular, the way context facilitates and influences the rise and flow of information." [Devlin, 2006]

Situation Theory reflects some peculiar ontological commitments with respect to the nature of information and how information *arises from* and *flows in* the world. Core to this ontological perspective is the notion of a *situation* and *infon*.

The notion of *situation* was first introduced by Barwise and Perry in [Barwise and Perry, 1980]:

> "The world consists not just of objects, or of objects, properties and relations, but of objects having properties and standing in relations to one another. And

> there are parts of the world, clearly recognized (although not precisely individuated) in common sense and human language. These parts of the world are called situations. Events and episodes are situations in time, scenes are visually perceived situations, changes are sequences of situations, and facts are situations enriched (or polluted) by language." [Barwise and Perry, 1980]

The notion of a *situation* has been defined in the Situation Theory literature in various, still related, ways where the word 'situation' may be understood in its normal, everyday sense, to refer to some part of the activity of the world.

Devlin describes situations as "part[s] of the activity of the world" (1991a, 11) or "highly structured [...] parts of the world that [an] agent's behavior discriminates". Gawron and Peters (1990a, 16) introduce situations as "limited parts of the world containing individuals and other objects, having properties and standing in relations". Ginzburg and Sag (2001, 83) conceive situations as "partial, temporally located, actual entities, whose role is to explicate such objects as states or events".[2]

Exactly what does and does not constitute a situation, according to Devlin, is largely a matter of the capabilities of a cognitive agent in her engagement with the world.

Situation Theory is mostly concerned with human agents but can be applied effectively to other types of agents, for example computing machines.[3]

Situation are *behaviourally discriminated* or *cognitively individuated* by an agent according to a *scheme of individuation.*

A *scheme of identification* is the way an agent "carves up reality" into cognitive "manageable pieces" [Devlin, 1995]. An agent recognises objects and other kinds of 'uniformities' in the world according to a view. It is the geometry, or syntactic rules, that allows an agent to distinguish *objects as objects* bearing some unity and identity criteria.

[2]These definitions, among others, have been collected by Jacob Ian Lee and presented in his MS thesis 'Situation Theory: a survey'. See http://jacoblee.net/wp-content/uploads/2011/11/ThesisJacobLee.pdf

[3]In [Devlin, 1995], Devlin illustrates an example of application for a simple robot.

One of the assumptions in Situation Theory is that "the world does not operate randomly from one moment to the next [...] but rather exhibits a great deal of regularity" [Devlin, 1995]. These regularities or *uniformities* are "cognitively individuated, or simply behaviourally discriminated" [Devlin, 1995] by cognitive agents according to their *scheme of individuation*. Generally speaking, a scheme of individuation represents the *ontology* an agent applies to make sense of its sensory perception of the world.

As they "make their way through the world" [Devlin, 1995], cognitive agents are exposed and acquire all sort of information from the environment.

Information is understood here in the ordinary use of the term 'information' with emphasis on its semantic features. According to Situation Theory, information is always 'extracted' by attending to discrete parts of the world —what Situation Theory calls *situations*.

Devlin appeals to Dretske's distinction between *analog* and *digital* representation of information [Dretske, 1983] —i.e. the capability to go "from the infinite continuous" that comes with the agent "fractal–like" and fuzzy perception of the world, to the "finite and discrete" that allows agents to "carve up reality into cognitive manageable pieces" [Devlin, 1995].

To avoid confusion with the use of the term 'digital' in digital preservation, a different terminology is adopted here, one that appeals to the levels of abstraction of information representation as identified by BRM. On this account, we align *analogue representation* and *digital representation*, respectively, with the notion of *concrete representation* and *symbolic representation*.

Therefore, a cognitive agent is an agent capable to *abstract* from the concrete to the abstract level.

The process itself involves two stages: *perception* and *cognition*. The first stage is *perception*, where

> "the environment becomes directly accessible to the agent by way of some form of sensor (seeing, feeling, smelling, hearing, etc. [...])" [Dretske, 1983].

The second stage is *cognition*. According to Dretske, *cognition*

> "is the *conceptual* mobilization of incoming information, and this conceptual treatment is fundamentally a matter of ignoring differences (as irrelevant to an underlying sameness) [...]" [Dretske, 1983].

The process of *abstraction* appeals to this capability of ignoring irrelevant differences, allowing, for example, the classification of entities in the world according to *types*: for instance, the type *animal* (an abstract object) is assigned to all living beings that exemplify certain characteristics (those encoded by the type *animal*).

Central to the discourse presented here, the process of *abstraction* allows us to recognise abstract objects *as individuals* separate from the concrete entities realizing them. For instance, the recognition of the same *sentence* (an abstract object of *symbol structure* kind) being realized by different utterances (concrete representations in the form of patterned sounds).

The relationship between abstract and concrete allows cognitive agents to make inferences between types of situations, such that a situation of one type carries information about a situation of the other. This kind of inference is governed by what situation theory calls *constraints*. A constraint is, in some sense, the background knowledge the agent can apply in the cognitive process.

Examples of constraints are "natural laws, conventions, analytic rules, linguistic rules, empirical lawlike correspondences, or whatever" [Devlin, 1995]. The constraints an agent is 'attuned to or at least aware of' [Devlin, 1995, Devlin, 2006] affect the capability of the agent to establish such relationships, ultimately establishing *what information* the agent is presented with in a given situation.

The *scheme of individuation* and *constraints* an agent brings to bear in a cognitive process form the *interpretive frame* [Dubin et al., 2011] the agent applies to her perception of the world in order to be presented with some information.

For a cognitive agent, *information* is said to be "carried by" or to "arise from"a situation — that is to say that situations function as *concrete representations* for the information they carry.

What information actually arises from a situation is function of the agent's *interpretive frame*.[4]

Information in this sense cannot be "reduced to" or be mereologically considered a "part of" any representation of that information, and is assumed to be built up from discrete "items of information" known as *infons*. Infons are discrete in Dretske's sense of *digital information* [Dretske, 1983].

The basic logical form of *information* —and as such of an *infon*— is that of an object having or not having a certain property. The general logical form of an infon is:

$$\ll R, a_1, \ldots, a_n, 1 \gg \textit{ or } \ll R, a_1, \ldots, a_n, 0 \gg$$

where R is an n–place relation and $a_1, \ldots, a_n$ are objects appropriate for R. What kinds of objects can fill the $a_1, \ldots, a_n$ roles is a function of

1. how an agent "carves out" the world into "manageable pieces" —i.e. her *scheme of identification*;

2. the type of the relation R.

Situation Theory also provides a mathematical construct called *abstract situation* to model *real situations* within its mathematical framework.

[4]An example from Devlin's book (Originally presented by Barwise in [Barwise, 1986]) makes vivid the contingent relationship between a situation, some information, and the constraint an agent is attuned to. Imagine a person coming across a tree stump in a forest —a *situation*. If she is "aware of the relationship between the number of rings in a tree trunk and the age of the tree [a natural law *constraint*] the stump can provide her with the age of the tree when it fell [certain *information*]" [Devlin, 1995]. If she is not aware of the relationship between the number of rings in a tree trunk and the age of the tree, she cannot obtain the information that the tree was of a particular age when it fell. However, if she is attuned to *other constraints*, she can obtain *other information* from the *same situation* —e.g. if she recognises the kinds of bark of the tree, then "the stump can provide the information as of what type of tree it was, its probable height, shape, leaf pattern, and so on" [Devlin, 1995].

> "There is an intuitive sense in which, to every real situation s, there corresponds a particular abstract situation, namely the set $\{\sigma | s \models \sigma\}$" [Devlin, 1995].

An *abstract situation* is a (potentially partial) model of a real situation. It is a *state of affairs* modeled as the set of *infons*.

By being partial models, abstract situations give the freedom to capture only aspects relevant to the particular analysis and facilitate "extensive mathematical modeling" [Devlin, 2006].

The freedom of construing abstract situation out of any set of infon can in fact lead to abstract situations that do not correspond to any situation in the real world. From an ontological perspective, *abstract situations* —being logical objects— describe *state of affairs* that might, or might not, *obtain* through a concrete situation.

Devlin's work provides both fundamental concepts, such as the notion of a situation, essential to the analysis presented here, and the cognitive–based and agent–driven approach towards "information flow" that grounds how we define requirements for successful communication.

2.4 Fundamental primitive kinds

A preliminary step for a throughout conceptual analysis of a domain —e.g. the digital preservation domain— is individuating foundational categories of things that will help consistently describing the domain of interest.

We provide therefore a simple set of foundational primitives useful to proceed with our conceptual analysis of information preservation. This set is based on the fundamental kinds identified within the Basic Representation Model (BRM) [Wickett et al., 2012] and complemented with additions from the Situation Theory ontology presented by Keith Devlin in his book "Logic and Information" [Devlin, 1995].

In this context, we consider an *object* to be any discrete individual in the world, one having certain *unity* and *identity* criteria that allow for its *discrimination*, possible *individuation*, and *reference* among other individuals.[5] When not specified otherwise, we will use the terms "object" and "entity" interchangeably.

In order to ensure consistency, we provide a simple set of fundamental *kinds of objects* that represent traditional categorical distinctions commonly applied in formal ontology and common sense discourse. The inclusion or otherwise of a particular kind in the ontology is functional to the modeling scope —i.e. provide a conceptual framework to effectively describe how information is *represented*, *communicated* and possibly *preserved*. Therefore, no commitment is taken on the 'true nature of reality'.

At the top level we distinguish between three *kinds of concrete objects* and three *kinds of abstract objects*.

2.4.1 Concrete objects

The three kinds of concrete objects are: *material object*, *event*, and *situation*.

Material object Something is a *material object* if it is an amount of patterned matter and energy [Sandore and Unsworth, 2010, Wickett et al., 2012]. Examples of material objects are physical hard drives and other hardware components. Material objects are typically understood as being wholly present at any time they exist [Casati and Varzi, 2010].

Event Something is an *event* if it occurs through time —i.e. it takes up time and persists by having different parts (or "stages") at different times. Examples of concrete events are: a particular football game or a particular uttering of a sentence.

[5] For more on the notion of *object*, see the entry 'Object' in the Stanford Encyclopedia of Philosophy at http://plato.stanford.edu/entries/object/

Situation Something is a *situation* if it is a temporally and spatially located structured part of the world. Situations involve material objects, events, and relationships among them, but any given situation is more then the sum of the objects, events, and relationships involved.

2.4.2 Abstract objects

The three kinds of abstract objects are *propositional content*, *symbol structure*, and *state of affairs*:

Propositional Content Something is a *proposition* if it is possible to believe or doubt it. A proposition is also typically understood as the language independent meaning of a sentence and the bearer of truth value.

Symbol Structure Something is a *symbol structure* if it is an abstract arrangement of symbols (individual symbol being an atomic symbol structure). Examples of abstract objects that can serve as symbol structures include graphs, relations, and sequences, along with more familiar kinds of symbol structures like strings of characters or sentences.[6]

State of Affairs Something is state of affairs if it is a way things might be. State of affairs in this sense closely relate to the notion of *proposition*: the sentence 'the cat is on his mat' is not only said to express the proposition that *the cat is on his mat*, but also to *describe* the state of affairs of *the cat being on his mat*. In fact, there are similarities between states of affairs and propositions. Propositions are true or false; states of affairs obtain or not.

These primitives will be referenced and utilized throughout the analysis that follows, in particular in Chapter 4, 5, 6, and 7, and be reiterated when appropriate.

[6]See [Wickett et al., 2012].

CHAPTER 3

THE CONCEPTUAL ANALYSIS OF *PRESERVATION*

3.1 Introduction

The notion of *preserving digital information* is fundamental in the broad digital and data stewardship agendas [Task Force on Archiving of Digital Information et al., 1996,Duerr et al., 2004]. Definitions of digital preservation and digital curation directly appeal to this idea. Expressions like "preserving digital information" [Hedstrom, 1997,Waugh et al., 2000,Chen, 2001] or "maintaining digital information" [CCSDS, 2002, Webb, 2003] are routinely used in the literature and scholarly discourse to refer to key objectives in the stewardship of our scientific and cultural digital heritage.[1]

Despite wide acceptance and routine use, the notion of *preservation* is still problematic in its application to digital artefacts. Perhaps this is not surprising.

What is surprising though, is that there has been so little effort to provide this notion, both fundamental and problematic, with a precise and rigorous formal definition.

In this chapter we will demonstrate that the lack of a formal analysis of digital preservation is indeed problematic, and this is primarily because the apparent parallels with physical preservation are misleading. Paradigms of traditional 'physical' preservation fail to deliver an appropriate account of what preservation really means in a digital context. Although some attempt has been made within the digital stewardship community to provide more adequate accounts of what *digital preservation* really entails [Ross, 2012]— misleading metaphors still

[1] See also [Yakel, 2007].

pervade the scholarly discourse in digital preservation.

What follows in this chapter are first steps towards a conceptual analysis of the notion of preservation, its relation with *information*, and the issues in its application in a digital context. We show that the traditional notion of *material preservation* only metaphorically applies to *information* and, consequently, cannot guide us at all in an analysis of digital preservation. In addition, the same strategy is adopted to demonstrate that, even apart from conceptual difficulties, preservation is not the best lens for characterizing the requirements for long–term digital stewardship.

3.2 What 'preservation' does mean

Dictionaries define 'preservation' as the "process of preserving something" and "the state of being preserved". The term derives from the verb 'to preserve',[2] defined as:

> A1. "maintain (something) in its original or existing state", A2. "retain (a condition or state of affairs)", A3. "keep from harm or injuries" (Oxford dictionaries[3])

Both senses of preservation —the former sense of emphasising an intentional process and the latter emphasizing the results of such process— are addressed in the definitions of *preservation* developed within the library and archival communities.

Dictionary definitions seem to imply the following account of material preservation:

MP0 x is preserved from time t_1 and time t_2 if and only if x exists at time t_1 and x exists at time t_2

This account, however, does not capture certain intuitions with respect to preservation practices: it is not enough for the material thing to *exist*. The thing undergoing *preservation*

[2]The verb "to preserve" originated from the Latin verb "praeservare": from prae- 'before, in advance' + servare 'to keep'.

[3]Entry "preserve" at http://oxforddictionaries.com/definition/english/preserve.

should also not loose certain properties, the ones that are considered essential for its identity as a particular object and are required to support its specific function.

If we look at definitions of *preservation* developed within the library and archives community, this perspective consistently emerges.

3.2.1 Definitions of preservation: the *material* paradigm

The first two definitions refer and apply to physical objects traditionally *preserved* in libraries and archives. For these reasons, we call them *material definitions*:

> [**D1**] "1. The professional discipline of protecting materials by minimizing chemical and physical deterioration and damage to minimize the loss of information and to extend the life of cultural property - 2. The act of keeping from harm, injury, decay, or destruction, especially through noninvasive treatment." (Glossary of Archival and Records Terminology [Pearce-Moses, 2012])

or

> [**D2**] "The overall package of administrative and/or practical measures, such as boxing, good housekeeping, careful handling and environmental control, which ensure the survival of documents without specialist intervention. Conservation and restoration procedures are part of a preservation policy." (UNESCO [Boston, 1998])

These definitions of preservation imply the existence of a *material object*. Successful preservation is described as "minimiz[ing] the loss of information and extending the life of cultural properties" and "survival of documents". In order to achieve successful preservation, "practical measures" must be established to ensure the protection of the material against potential decay. There is a strong alignment between these definitions and the dictionary definitions of the term 'preservation'. Preserving an object in the literal sense of *preservation*

seems to imply that an object is preserved when it persists through time without relevant changes in its significant characteristics, where significance is defined with respect to a particular community of stakeholders against which successful presentation is assessed.

It also implies —at least when the verb 'to preserve' is used in the active voice— that the object persists not by chance, but because procedures have been established to ensure and assess its persistence.

A formal definition of this notion might be:

MP1 A physical object x is preserved from time t_1 to time t_2, relative to stakeholder community C_S, if and only if for all ϕ, if ϕ is a property of x that is significant for C_S, and x has ϕ at t_1, than x has ϕ at t_2

This account of preservation certainly applies to material objects —objects that can change, be damaged, or decay— and focuses on the materiality of library and archival resources.

We make several observations about **MP1**. First, it relativizes preservation to particular communities that determine which properties of an object are significant or important and therefore should be retained by an object undergoing preservation. This means that an object may be preserved with respect to one community, but not another. Second, once a particular community is identified, preservation appears to be a binary determination — an object is either preserved or it isn't— and there is a high bar: all properties that are significant should be retained by a successfully preserved object. Finally, we note that this definition of preservation does not require *intent* to preserve.

This is a sense of *preservation* that obviously requires that a preserved object continues to exist (x has the same referent at t_1 and t_2). While it does not explicitly allow that an object can fail to be preserved and yet continue to exist, that is nevertheless a plausible background assumption.

The *information* library and archive materials carry (or are supposed to carry) does not directly enter into the equation at this level: there appears to be an implicit assumption that

preserved objects can, all other things being equal, fulfill their intended *informing* functions.

3.2.2 Definitions of preservation: the *transitional* paradigm

Whereas the *material* accounts of preservation foreground what is constant over time, the definitions provided by the Online Dictionary of Library and Information Science and InterPARES present a very interesting shift in terminology that foregrounds *intentional change* as an intrinsic part of preservation, accommodating the practice of media conversion and reformatting. For this reason, we call them *transitional*.

> **[D3]** "Prolonging the existence of library and archival materials by maintaining them in a condition suitable for use, either in their original format or in a form more durable, through retention under proper environmental conditions or actions taken after a book or collection has been damaged to prevent further deterioration." (ODLIS [Reitz, 2004b])

and

> **[D4]** "The whole of the principles, policies, rules and strategies aimed at prolonging the existence of an object by maintaining it in a condition suitable for use, either in its original format or in a more persistent format, while leaving intact the object's intellectual form." (InterPARES 2 Terminology Database [InterPARES, 2013])

The ODLIS definition, while suggesting traditional means intended to preserve a (physical) object against change (e.g. damage or decay), allows for intentional changes in the "format" or "form" of an object when a new form would be "more durable".

The InterPARES definition shifts even more from the previous ones. The goal here is to prolong the existence of an *object* (of unspecified type) by maintaining *it* "in its original

format[4] or in a more persistent format, while leaving intact the object's *intellectual form*"[5] [InterPARES, 2013].

Although the accounts presented in the ODLIS and InterPARES definitions are familiar and perhaps, at some level, even plausible, it is difficult to understand them literally. What does it mean for an object to be in one format or another? What sort of thing is it, exactly, that can be preserved if moved from its original format to another format, but might not be preserved if this transformation is not made? What exactly is this "intellectual form" that might or might not remain "intact" across the transformation? InterPARES does provide a discussion but it is not immediately evident how to understand these concepts from a formal perspective.

Notwithstanding the unclear ontological status of such *objects* and *materials*, it appears that both the InterPARES [InterPARES, 2013] and ODLIS [Reitz, 2004b] definitions are clearly engaged with the hard problem of describing a different sort of preservation, one that moves beyond the preservation of the original physical artefact.

What exactly then does *preservation* really entail for InterPARES? And what kinds of things can be objects of *preservation* in this sense?

While the ontological status of the *material* [Reitz, 2004b] and *object* [InterPARES, 2013] mentioned above is unclear, we can attempt a provisional formalization. Assuming that the original format of something x to be preserved reflects a suitable condition for its use by a community of stakeholders C_S, we might say that:

TP1 An object x is preserved between time t_1 and time t_2 if and only if x has format y at t_1, x has format z at t_2 and y is *intellectual-form-equivalent* to z with respect to x.

As well as the problem of defining "format", "intellectual form", and *intellectual-form-equivalent* for objects, there is the related problem of determining what sort of objects these

[4]Italics by the author of this dissertation.
[5]Italics by the author of this dissertation.

are that are being preserved via the management of equivalent formats. It seems unlikely that all of these objects are particular material objects. Potential candidate categories might be abstract objects such as *content* or *information* —or, alternatively, abstract *representations* of some content or information, such as a text— where change is admitted in the physical carrier that realizes such abstract object. *Has format*, on this account, is intended as a relationship between the abstract object to be preserved (e.g. some information) and a particular material carrier (e.g. a paper document).

Despite the potential variation in the *level of abstraction*, a possible approach in this sense is to situate our x (at least for library materials) at one of the levels described by the FRBR Group 1 entities. Depending on the preservation intent and expectations, and given a material carrier (an individual situated at the FRBR Item level), x might be understood as a) a particular *work*; b) an *expression* of such a *work* (such as a particular *text*); or even c) a *manifestation* (when presentational aspects are considered).

The approach we are taking here, appealing to the relational notion of equivalence, also reflects a different, yet related, approach to understand (and describe) library material. Elaine Svenonious, in her influential book *The intellectual foundations of information organization* [Svenonius, 2000], adopts a set theoretic approach where the notion of *equivalence* plays a central role in understanding how library materials can be effectively described and organized.

Even when digital preservation is not about preserving the digital counterpart of library material, related questions regarding levels of abstraction arise within the digital preservation paradigm discussed next, where the notions of *information* and *content* explicitly appear.

3.2.3 Definitions of preservation: the *digital* paradigm

Consider the following definitions of *digital preservation*:

> **[D5]** "The process of maintaining, in a condition suitable for use, materials

produced in digital formats, including preservation of the bit stream and the continued ability to render or display the content represented by the bit stream." (ODLIS [Reitz, 2004a])

[**D6**] "The act of maintaining information, in a correct and independently Understandable form, over the Long Term." (OAIS Glossary of terms [CCSDS, 2002])

[**D7**] "The active management of digital content over time to ensure ongoing access." (Library of Congress NDIIPP program)

[**D8**] "The series of managed activities necessary to ensure continued access to digital materials for as long as necessary. [...] all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change." (Digital Preservation Coalition)

While *preservation* continues to be the *definiendum*, when we compare these definitions to the *material* and *transitional* ones, the shift in terminology and concepts is quite vivid: *digital preservation* is now about *ensuring access* to *information*, *content*, and *(digital) materials*, not about ensuring the existence or integrity of a physical object.

Despite this shift, the term 'preservation' also continues to appear in expressions like 'preserving a digital object' or 'preserving digital information' that are widely adopted within the digital preservation literature.

Unfortunately, while the *material* and, to a certain extent, *transitional* definitions of preservation directly reflect key features of our shared understanding of the term and give a relatively simple account of what successful preservation entails, the definitions of *digital preservation* presented above appeal to notions that do not provide a familiar, let alone precise, account of what digital preservation is, or how to recognize success.

What do we mean by "maintaining, in a condition suitable for use, materials produced in digital formats"? What is it to "maintain access to digital materials"? What about "maintaining information, in a correct and Independently Understandable form"?

These kinds of definitions do not support any immediate attempt to derive a more formal account of what preservation means in a digital context. To see this, note for example that the concepts in **MP1** are relatively simple. How could a comparably precise formal definition be provided using concepts like *maintaining* or *access*? Moreover, these definitions do not provide an account of what *digital material* or *information* are, accounts that are necessary in order to identify successful preservation.

Let's consider then what kinds of things can be, and are, *literally* preserved in a digital preservation context.

3.3 Conclusion

Does *preservation* really apply in a digital context? Although there are well–known puzzles regarding the identity of physical objects [Haslanger and Kurtz, 2006], there is no doubt that the notion of *preservation* has certainly rather well understood implications with respect to material objects, and assumes that those objects can survive not only losing and gaining properties, but also the rearrangement and even loss of physical parts. *Preservation* here should be distinguished from *conservation*: "conservation counters existing damage, as distinguished from preservation, which attempts to prevent damage".[6]

Physical objects can be damaged, decay, and, more generally, are by their nature prone to change over time. Therefore, when the targets of preservation are physical objects —such as the traditional physical library and archival materials collected within the walls of library and archive facilities— the action of preservation may be applied and is, indeed, required in order to prolong their existence through time. Definitions of preservation following the

[6]See http://www2.archivists.org/glossary/terms/c/conservation.

material paradigm certainly reflect this notion.

However, as recognised early on by the digital preservation community, the continuity of physical components does not seem to be relevant, at least in the same way, to the identity of *digital objects*.[7]

While a hardware infrastructure is required to access and interact with digital materials, no particular hardware component (e.g. a particular hard drive) is typically understood to be an essential component of any digital material. This is reinforced by the common assumption that *the same* digital material (e.g. a digital document, an e-book, a web page etc.) can be multiply and repeatedly 'copied' over different media devices. Even considering cases where the bits cannot be migrated to new physical media — e.g. video game cartridges where the source is not open— this is a contingent fact, not an essential characteristic of the digital material itself.

In fact, digital preservation efforts many times involve the deliberate migration to new physical media: digital preservation is "plagued by the short [physical] media life" [Chen, 2001] because "storage media, such as disks, tapes and cartridges, decay relatively rapidly compared to other media" [Heslop et al., 2002] that have been traditionally used to carry information (e.g. paper, parchment, papyrus, etc.).

If material objects involved in access, use, and interaction with digital objects do not directly participate in their identity —it is common intuition that *same* digital object can be replicated on different storage media and be accessed in different technology environments— then what is it that must be 'preserved' or 'maintained' in order to 'ensure continuous access to digital materials'?

An answer typically considered legitimate, if not complete, is that we need to *preserve the bits* or *preserve information*. These two perspectives are discussed in the following chapter.

[7]The term 'digital object' is used at this point in its colloquial underspecified sense. No commitment over its ontological kind is assumed, other than the intuitive notion that digital objects are not physical objects.

CHAPTER 4

PRESERVING THE *BITS* AND PRESERVING *INFORMATION*

4.1 Introduction

We saw in the previous chapter how definitions of *preservation* evolved to fit evolving concepts in the preservation community, beginning with the naïve version of material preservation and proceeding to the focus on maintaining access. Before going on to examine the sophisticated influential theories of digital preservation in Chapter 5, we should take a closer look at the notion of *preserving the bits* and the notion of *preserving information*. Both of these concepts continue to play significant roles in the preservation discourse.

4.2 What 'preserving the *bits*' does mean

The Library of Congress National Digital Stewardship Alliance considers bit preservation the first tier of its *preservation levels*:

> as one moves up each of the tiers from Level 1 to Level 4, one is moving from the basic need to ensure bit preservation towards broader requirements for keeping track of digital content and being able to ensure that it can be made available over longer periods of time.[1]

Consider the need to preserve a recorded bit sequence of interest, and represent that bit sequence in new media with some warrant of authenticity and integrity [Giaretta et al., 2009]

[1] See http://www.digitalpreservation.gov/ndsa/.

in some sense of those concepts. Even here, at this relatively low 'physical' level, we already encounter conflict with the metaphor of preserving the relevant characteristics of a physical object: there is no (literal) sense in speaking of 'preserving' a bit sequence as if it were the sort of object that is subject to corruption. The particular bit sequence that is involved in the process is essentially a sequence of 1s and 0s,[2] an *abstract symbol structure* that can be repeatedly and multiply instantiated in various physical media.

If properly performed, copying a DVD–ROM —i.e. "burning" a new DVD–ROM with the same relevant physical characteristics as the original one— creates a completely *new inscription* or *new instantiation* of exactly the *same (set of) bit sequence(s).* The persistence of bit sequences is ensured by their fundamental nature as repeatable abstracta (rather than concrete physical objects): they are objects that are not created and cannot change, or be damaged or destroyed. Therefore preservation, in its literal sense, is neither required nor possible.

The expression 'preserving the bits' appears to be a linguistic convention shorthanding a more complex set of assumptions: bit preservation implies preserving an analog arrangement of physical material from which a computer can derive a signal of some sort that by relevant *conventions* is given a digital *interpretation.*

4.2.1 Preserving the bits: an account

Following the premise above, we can provide an account of *bit preservation* that does not appeal to the literal sense of the term 'preservation'.

At a high level of generality, *bit preservation* means enabling the *possibility* for the *same (set of) bit sequence(s)* to be *discriminated* at different points in time, and, potentially, across changes in the underlying storage technology.

[2]More precisely, a bit sequence is an abstract mathematical object: it is a *sequence*, a function having as domain the set of natural numbers and as codomain the set {0,1}.

By *discrimination* we mean here that, in virtue of certain physical arrangements in the machine, some abstract patterns are recognized.

On this account, preservation involves an active discrimination process that can only happen at run–time when a digital environment (DE) is involved. From a high–level, a DE can be understood as a complexion involving a set of interconnected hardware and software components.[3] At run time a DE takes up an agentive role (called here *machine agent*) with respect to the processes happening within DE. These processes are enabled when, at run time, software components become available providing the rules against which a DE operates.

Certain configurations of hardware and software components within a DE are recognized at run time as performing a *persistent storage* function.

Given sets of interpretive rules IF_1 and IF_2 (i.e. two *interpretive frames* [Dubin et al., 2011]) applied in the process, we can derive the following non–literal account of bit preservation (BP), where bit preservation is understood as *possible bit discrimination*:

> **BP** A bit sequence x is *preserved* between time t_1 and time t_2 if and only if there exists a DE such that:
>
> 1. there is a persistent storage ps_1 recognised as part of DE at time t_1 such that it is possible for DE to discriminate x from y when IF_1 is applied;
> 2. there is persistent storage ps_2 part of DE at time t_2 such that it is possible for DE to discriminate x from z when IF_2 is applied;
> 3. (2) happens because ps_1 is intended to *realize* x in DE according to IF_1 at t_1 and ps_2 is intended to *realize* the same x in DE according to IF_2 at t_2.

No assumption here is made on whether ps_1 and ps_2 are different persistent storage.

[3]No assumption here is made on the way hardware and software components are related and connected: a DE can be in this sense a local machine, but also a distributed cloud service, a regional cluster, or a client/server architecture where hardware and software components interoperate via network connections.

The *intentionality* in clause (3) in intended to capture an *active preservation intent*: x can be recognised in DE at time t_2 not *by chance* but because established procedures are in place: ps_2 is physically arranged in such a way that is 'bit-equivalent' to ps_2 with respect to x when, respectively, IF_2 and IF_1 are applied in the discrimination process.

The *possibility* in clauses (1) and (2) is not *logical possibility* nor *metaphysical* or *nomological possibility*, but a more colloquial sense of possibility connected with the capability of technical environments.

This account of preservation as *potential discrimination* is by definition *potential*. *Actual preservation* can only be assessed at run time: the *possibility* becomes *actuality* only when the intended bit sequence x is discriminated within DE.

It is worth noting that there is no certainty of successful *bit preservation* other than *in the moment* when the discrimination process happens. This implies that between time t_1 (the time when x entered preservation) and a later time t_2 (the time when preservation is assessed) we can only *claim* that procedures have been established (if any) that lower the threats for an actual future discrimination of the same bit sequence x.

It should not come as a surprise that, in the latest years, management practices in digital preservation have drawn heavily from risk management as a methodology [Lee et al., 2002, Strodl et al., 2007, Becker et al., 2009, Barateiro et al., 2010, Ross, 2012]: we are not dealing with *certainty* but with *possibility*.

4.2.2 Does preservation apply to the *bits*?

The lesson learned here is that there is no literal bit preservation. Preserving the bits is a metaphor that can only be explained in terms of contingent relationships between material objects (e.g. a persistent storage) and abstract objects (such as a bit sequence) with respect to an agent capable to properly recognise such relationship. From this perspective, routine auditing procedures —involving checks on successful bit discrimination— are the only

method to assess successful preservation over time.

Understanding *bit preservation* as *potential discrimination* also helps to explain recurring expressions such as 'bit rotting' or 'bit corruption'. The change (rotting, corruption, etc.) does not happen at the abstract bit sequence level but at the physical storage media level. Simply, the *is realized by* relationship between the intended bit sequence and a storage situation ceases to hold (for example in virtue of unintended physical changes in a storage media device) and a *different* bit sequence (or no bit sequence at all) is realized for the processing agent in place.

This basic perspective on bit preservation, however, does not present the complete picture of digital preservation challenges.

Bit level preservation is a *mean*, not the *goal*, in digital stewardship. As suggested by the OAIS definition of digital preservation, successful digital preservation is about "maintaining" or "preserving" *information*. The concept of *information preservation*, however, appears to be as challenging as *bit preservation*.

4.3 What 'preserving *information*' does mean

Understanding what *preserving information* means requires some preliminary observations on the notion of *information* itself.

4.3.1 A preliminary perspective on *information*

The concept of *information* has been the object of extensive study in the field of Information Science. Yet, there is still no agreement on what information really is. The issue has been addressed from many different perspectives, ranging from the more 'practical' or 'operational' —where the concept of information is only appealed to and not necessarily analytically described— to the more theoretical ones —where information is attempted to be precisely

described as a *kind of thing*. Comprehensive studies have tried to capture and review the different perspectives and approaches in a systematic way (e.g. [Capurro and Hjørland, 2005]).

Despite the various accounts of what information might be, there are practical reasons to position the notion of information at the abstract propositional level, following a strategy similar to those adopted to explain related concepts like *meaning* and *knowledge*.

This strategy defines the notion of information in relation with scenarios where an agent *is informed that* p, and where p is the information content that the agent is presented with as a result.

Consider the following examples: Claudia, a fictional character, participates in ordinary situations where information, in its intuitive sense, appears to be involved. These examples shows how Claudia obtains the same information in different ways.

A Claudia engages a Chicago Transportation Authority (CTA) officer in a conversation to obtain information about buses to downtown. When Claudia asks the officer, the officer informs Claudia that *the bus 147 to downtown leaves at 4:15PM* (some information);

B Claudia picks at the bus station a CTA bus schedule printed brochure. By 'reading' the brochure, Claudia is informed that *the bus 147 to downtown leaves at 4:15PM* (some information);

C Claudia 'opens' a digital version (e.g. a digital brochure in PDF format) of the CTA bus schedule on her computer. By 'reading' the digital brochure on screen Claudia is informed that *the bus 147 to downtown leaves at 4:15PM* (some information).

In these examples, Claudia engages with the external world with her senses, and from her *experience* of the world she comes to stand in a particular relationship with something —namely that *the bus 147 to downtown leaves at 4:15PM*— that is *about* part of the world. This *something*, we name it p, is what we colloquially call *information*.

Intuitively p is something that can be believed or doubted and that it is either true or false. It also a thing that is *independent* of the specific *informing situation* Claudia might be involved in: Claudia can be presented with the same p under different circumstances where p is *represented* and *communicated* differently.

For these reasons, we position the notion of information, ontologically, at the *propositional content* level.

However, it is not the case that any propositional p is information *per se*, and not every proposition 'in the wild' is information. A propositional p becomes information when participating in an *informing situation*, where an agent, such as Claudia, is *informed* that p. On this account, *being information* is a contingent role that p might play, but *being information* is not an essential property of any p.

Information can then be defined as follows:

Definition 1 (Information) *p is information at time t_1 if and only if p is a proposition and there is an agent a (such as a person) that is informed that p at time t_1.*

A proposition is defined in the classical way:

Definition 2 (Proposition) *p is a proposition $\stackrel{\text{def}}{=}$ it is possible for someone to believe p.*

Information–as–proposition is therefore an object of abstract sort.

In order for an agent to be presented with a particular p there is always an *interpretation process* in place where specific rules and conventions are applied. This process always involves

objects that are concrete —and therefore can be object of sensory perception— that function as *representations* for that p.

In our examples, a *spoken utterance* (Example A), a *material object* (Example B), and an *output on screen* (Example C) function as different *representations* of the same p.

A similar perspective on information is the foundation of one of the most influential models developed within the digital preservation community: the Open Archival Information System (OAIS) Reference Model [CCSDS, 2002].

The OAIS *Information Model* is the specific component of the OAIS Reference Model that models the requirements for information access and preservation.

Following, we discuss the OAIS Information Model and provide some observations about its insights and issues in the lens of this account of information. From these observations we will draw some preliminary conclusion on the notion of *information preservation*.

4.3.2 The OAIS Information Model

Defined by recommendation CCSDS 650.0-B-1 of the Consultative Committee for Space Data Systems [CCSDS, 2002] and corresponding to ISO 14721:2003, the OAIS Reference Model directly engages with the problem of relating *information* with the bits via its *Information Model*.

The OAIS *Information Model* is the specific component of the OAIS Reference Model that models how information is "handled by an OAIS" [CCSDS, 2002, Page 4-1], the assumption being

> "Data interpreted using its Representation Information yields Information" [CCSDS, 2002, Page 2-4]

where *information* is defined as:

> "Any type of knowledge that can be exchanged. In an exchange, it is represented

by data. An example is a string of bits (the data) accompanied by a description of how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius (the Representation Information)."

Accompanying the text is the diagram in Figure 4.1 that represents relevant entities and relationships:

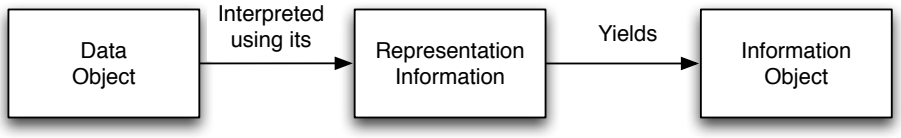

Figure 4.1: Obtaining Information from Data in OAIS

Data is modeled in OAIS as a *Data Object*:

"The Data Object may be expressed as either a physical object (e.g., a moon rock) together with some Representation Information, or it may be expressed as a digital object (i.e., a sequence of bits) together with the Representation Information giving meaning to those bits." [CCSDS, 2002, Page 4-20]

An *Information Object* is defined as:

"A Data Object together with its Representation Information" [CCSDS, 2002, Page 1-12]

where *Representation Information* is intended as:

"The information that maps a Data Object into more meaningful concepts." [CCSDS, 2002, Page 1-14]

Preserving a *data object* is necessary but not sufficient: we need to preserve also the interpretive means —i.e. the *Representation Information*— such that the user of a *designated*

community can properly access information in a meaningful way.[4] Representation Information in this sense constitutes the complex set of *interpretive frames* [Dubin et al., 2011] that, comprehensively, maps the bits to the information a user should be presented with.

Consider the following examples from OAIS:

> "An example of Representation Information for a bit sequence which is a FITS file might consist of the FITS standard which defines the format plus a dictionary which defines the meaning in the file of keywords which are not part of the standard." [CCSDS, 2002, Page 1-14]
>
> or
>
> "Another example is JPEG software which is used to render a JPEG file; rendering the JPEG file as bits is not very meaningful to humans but the software, which embodies an understanding of the JPEG standard, maps the bits into pixels which can then be rendered as an image for human viewing." [CCSDS, 2002, Page 1-14]

These examples make vivid the OAIS approach, in particular:

1. information is "always expressed (i.e. represented) by some type of data" [CCSDS, 2002, Page 2.3], which is to say that information can only be accessed through a *representation* of that information.

2. in order for a user to be presented with the *intended information*, an appropriate interpretation of the data needs to happen, and such an interpretation is always governed by an interpretive frame —a specific set of rules and conventions that map the intended information to a representation. Such rules and conventions are modeled in

[4]The OAIS recognises that we should also consider the background knowledge of the designated community and the representation information that needs to be preserved is complement of such background knowledge. On the topic see also [Giaretta, 2008, Chowdhury, 2010, Strodl et al., 2007, Becker et al., 2009].

OAIS as a combination of Representation Information and the background knowledge of a designated community for which information should be 'preserved'.

3. OAIS emphasises that successful preservation can only be assessed with respect to a designated community of users: their capability and expectation need to fit into the equation.

4.3.3 Does preservation really apply to information?

If information is essentially an entity of abstract propositional sort —and certainly OAIS seems to agree when defines information as "Any knowledge that can be exchanged" [CCSDS, 2002]— preservation in its literal sense does not apply to information, nor it is required. Similarly to bit preservation, we are facing a situation where the object to be preserved cannot be damaged, destroyed, or decay. Similarly also, there is an interpretation process that needs to happen.

Preserving information appears to be, in fact, a metaphorical expression, a sort of shortcut, where a complex set of requirements needs to be satisfied in order for an agent to be presented with an intended information p.

Interestingly, however, at least from the user perspective, bits are not directly involved in this interpretation process —as suggested by Example C presented above where Claudia engages with a digital brochure through an output on screen.

While, to a certain extent, we can understand the bits — a *Data Object* in OAIS— as ultimately representing the information that a digital object is intended to carry for a user, they do not express such information *directly*. In most cases,[5] bit level representations do not constitute meaningful representations at the user level —i.e. a bit level representation is not the representation intended for user access, its "documentary form" [Duranti, 2005].

Bits need to be appropriately *interpreted* in order to provide access to objects of other

[5]Unless the object of analysis for a user *is* a particular bit sequences.

sorts, and are these objects that are meaningful for human agents. Objects of the latter sort are typically understood either as the digital counterpart of well understood information–carrying material objects —where 'digital' is typically used as a prefix such as in *digital document*, or *digital picture*— or other objects only recognized in the digital domain —e.g. *web page*, *video game*, etc.

Terms like 'digital document', 'e–book', or 'web page' —or general terms such at 'digital material' hardly denote objects at the bit sequence level and their reference, if any should be accounted for.

4.4 Conclusion

In this chapter we explored some plausible approaches to concepts of *preservation* and we showed that preservation, in its literal sense, does not apply to *bit sequences* and *information*, entities typically understood at targets of digital preservation efforts.

We also showed that the influential OAIS Reference Model [CCSDS, 2002] reflects these conclusions, and adds further critical insights on the relationships between bit level representations and information: *access to information* can only happen when a *Data Object* is properly *interpreted* according to appropriate *Representation Information*.

OAIS, however, does not capture all the entities that appear to be involved in this interpretation process. In particular, OAIS fails to capture the sort of entities users are engaged with when they interact with digital materials, those entities that actually represent information from the user perspective.

In the next chapter we discuss four different, yet related, conceptual perspectives that attempt strategies to address these issues, drawing some conclusions about the actual nature of digital materials and their function as information carriers.

CHAPTER 5

PRESERVING *ACCESS* TO DIGITAL MATERIALS

5.1 Introduction

In the last chapter we explored some plausible approaches to a theory of preservation, noting difficulties with naïve accounts and laying the groundwork for an improved approach.

We have seen that, not only digital preservation is not bit preservation, but that bit corruption is, strictly speaking, not possible. Similarly, we have seen that information is also safe from corruption, and so it is hardly the proper object of preservation. Indeed, the best contemporary theories of digital preservation do not focus on the preservation of any sort of object, but rather on preserving *access*. But understanding what 'preserving access' means —and what is accessed and how— presents deep conceptual challenges.

In this chapter we explore in depth three influential theories of preservation.

5.2 Three approaches to digital preservation

Experiencing digital materials requires a combination of hardware and software: their access is "mediated by technology" [Heslop et al., 2002].

> "If the software for making sense of the bits (that is for retrieving, displaying, or printing) is not available, then the information will be, for all practical purposes, lost." [Kuny, 1998]

The relationship between bits and other objects that are useful and effective to characterize the complex nature of digital objects has been extensively analyzed within the digital preservation community. In the section we will focus on three conceptual approaches that we consider extremely influential for their many insights: Thibodeau's concept of a *digital object* [Thibodeau, 2002], the InterPARES approach towards understanding digital records, and the National Archives of Australia Performance Model [Heslop et al., 2002].

These works try to capture, beyond bit sequences, complementary aspects of what needs to happen in a digital environment in order for *users* to access digital material in the form "purportedly intended for user consumption" [Duranti, 2005].

5.2.1 Thibodeau's concept of a digital object

Kenneth Thibodeau presented his concept of a digital object in a 2002 influential paper titled "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years" [Thibodeau, 2002].

According to Thibodeau, a *digital object* can only be explained in terms of other objects:

> "All digital objects are entities with multiple inheritance; that is, the properties of any digital object are inherited from three classes. Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly different. A physical object is simply an inscription of signs on some physical medium. A logical object is an object that is recognized and processed by software. The conceptual object is the object as it is recognized and understood by a person, or in some cases recognized and processed by a computer application capable of executing business transactions."[1] [Thibodeau, 2002]

[1]Thibodeau's levels resemble closely the typical levels understood at the core of the notion of *data independence* in the theory and practice of database management systems.

The physical level represents the hardware layer in a digital environment, "inscription of signs on a [physical] medium":

> "Basically, the physical level deals with physical files that are identified and managed by some storage system. The physical inscription is independent of the meaning of the inscribed bits." [Thibodeau, 2002]

The logical level represents entities (such as abstract bit sequences, or higher level data structures) that are recognised and processed "according to the logic of a software application" [Thibodeau, 2002]:

> "Once data are read into memory, the type of medium and the way the data were inscribed on the medium are of no consequence. The rules that apply at the logical level determine how information is encoded in bits and how different encodings are translated to other formats; notably, how the input stream is transformed into the system's memory and output for presentation." [Thibodeau, 2002]

The conceptual level captures the "presentation layer" or the characteristics of objects such as those "we deal with in the real world" [Thibodeau, 2002]:

> "[A conceptual object] is an entity we would recognize as a meaningful unit of information, such as a book, a contract, a map, or a photograph. In the digital realm, a conceptual object may also be one recognized by a business application, that is, a computer application that executes business transactions." [Thibodeau, 2002]

Thibodeau's objects (conceptual, logical, physical) are FRBR-like and there is an almost endemic tendency throughout information organization to see property inheritance between

FRBR-like object classes. This is first identified in the Renear and Choi paper devoted to the topic [Renear and Choi, 2006].

Despite the unfortunate use of the notion of inheritance,[2] Thibodeau's account of a digital object provides many compelling insights into the nature of what we colloquially call *digital objects*, how we access them, and what their preservation implies.

His notion of *conceptual object* is particularly interesting because it closely resembles notions like *digital material* that repeatedly appear in definitions of digital preservation as those objects "recognized and understood by a person".

The following quotation summarized his perspective on the requirements for successful digital preservation:

> "The process of preserving digital objects is fundamentally different from that of preserving physical objects such as traditional books or documents on paper.

[2]According to Thibodeau, an instance of the class *digital object* inherits properties from three distinct classes: *Physical Object*, *Logical Object*, and *Conceptual Object*. This perspective presents some ontological challenges. Can we really conceive an individual object that is at the same time, for example, an inscription (in the physical sense of inscription) and the "data that is loaded into the memory" of a computing machine? This seems quite unlikely.

Multiple inheritance is a common concept in knowledge representation and applies to a class of individuals when the class is a subclass of (a subsumption relationship) multiple classes. As an example, consider the class of all platypus (See ¡http://en.wikipedia.org/wiki/Platypus¿). We can say that instances of the class *platypus inherit* their properties from both the class of all *mammals* and the class of all *egg-laying animals*. This means that a platypus instantiates all the characteristics associated essentially with both the class *mammal* and the class *egg-laying animal*.

These problems of adopting a multiple inheritance approach can be made vivid by formalizing the subsumption relationships in First Order Logic:

$$\begin{aligned} &\forall(x)(DigitalObject(x) > PhysicalObject(x)) \ \& \\ &\quad (DigitalObject(x) > LogicalObject(x)) \ \& \\ &\quad (DigitalObject(x) > ConceptualObject(x)) \end{aligned} \tag{5.1}$$

Any property ϕ of a superclass is inherited by each of its subclasses, therefore:

$$\begin{aligned} &\forall\phi, (\phi(PhysicalObject) > \phi(DigitalObject)) \ \& \\ &\quad (\phi(LogicalObject) > \phi(DigitalObject)) \ \& \\ &\quad (\phi(ConceptualObject) > \phi(DigitalObject)) \end{aligned} \tag{5.2}$$

Considering his description of the individual classes and their characteristics —and notwithstanding any interpretation of what a *digital object* really is— there no such individual thing that can possibly be at the same time an instance of the three classes *physical object*, *logical object*, and *conceptual object*.

> To access any digital object, we have to retrieve the stored data, reconstituting, if necessary, the logical components by extracting or combining the bit strings from physical files, reestablishing any relationships among logical components, interpreting any syntactic or presentation marks or codes, and outputting the object in a form appropriate for use by a person or a business application. Thus, it is impossible to preserve a digital document as a physical object. One can only preserve the ability to reproduce the document. Whatever exists in digital storage is not in the form that makes sense to a person or to a business application. The preservation of an information object in digital form is complete only when the object is successfully output. The real object is not so much retrieved as it is reproduced by processing the physical and logical components using software that recognizes and properly handles the files and data types." [Thibodeau, 2002]

With this addendum:

> "The process of digital preservation, then, is inseparable from accessing the object. You cannot prove that you have preserved the object until you have re–created it in some form that is appropriate for human use or for computer system applications." [Thibodeau, 2002]

While Thibodeau's rich narrative captures familiar intuitions about the challenges we face in digital preservation, it does not deliver a complete picture, leaving some foundational questions unanswered.

While a *physical object* and a *logical object* can be safely understood, respectively, as an object of concrete material kind and an object of abstract symbol structure kind, the notion of *conceptual object* is less clear, ontologically.

A conceptual object is described both as "an object we deal with in the real world" and an object that can be repeatedly "re–created" or "reproduced" in some "form that is appropriate for human use or for computer system applications" [Thibodeau, 2002].

This notion appears to share characteristics with *both* a *repeatable abstraction* that can be multiply instantiated and a tangible *material object* that a user can directly experience through her senses.

Both aspects should certainly be taken into consideration.

On one hand, a conceptual object appears to be something that is "encoded" in (or by) a logical object and that "can be used by machine application". Even without a precise understanding of the *encoding* relationship, we can hardly conceive such an object to be a material object.

On the other hand, a *conceptual object* appears to be something that users can *directly* recognise and engage with: a thing a user can see, hear, etc., something like a "a book, a contract, a map, or a photograph" [Thibodeau, 2002]. Thibodeau, however, certainly does not refer to familiar material artefacts such as a (physical) book or a (physical) photograph in his account, but to entities that share certain characteristics with their material counterpart.

What would be then this "*real* object [that] is not so much retrieved as it is reproduced by processing the physical and logical components" and that is "in some form that is appropriate for human use or for computer system applications" [Thibodeau, 2002]?

We are possibly dealing with a compound entity that requires further analysis to be precisely defined.

While Thibodeau's observation that "the process of digital preservation is inseparable from accessing the [conceptual] object" seems to suggest that a *preservation–as–access* account can be attempted, when we try to formalize such an account adopting his perspective, we fall short.

If a *conceptual object* is an abstraction, what would be the concrete persistent object that realizes it such that it is possible for a user to recognise it?

If a *conceptual object* is instead concrete, what sort of concrete entity is *temporary*, reproducible still being the same, and dependant upon certain activity happening in a machine?

The National Archives of Australia *Performance Model* engages with this issue in a compelling way by appealing to the notion of a *performance*.

5.2.2 The NAA Performance Model

The National Archives of Australia Performance Model, developed in the context of the Agency to Researcher Digital Preservation Project hosted at the National Archives Of Australia and presented in a 2002 report titled "An Approach to the Preservation of Digital Records" [Heslop et al., 2002], proposes a perspective that is apparently radically different from the ones we previously presented.

The rationale for the NAA Performance Model is that

> "Digital records challenge the idea that records are essentially objects for archivists to preserve, arrange, store and make accessible" [Heslop et al., 2002].

In particular:

> "[Digital records,] while fulfilling the same general business purpose [of paper records], are mediated by technology, which means that to experience digital records a person must have the right combination of hardware and software. [...] The experience of the object only lasts for as long as the technology and data interact. As a result, each viewing of a record is a new 'original copy' of itself - two people can view the same record on their computers at the same time and will experience equivalent 'performances' of that record." [Heslop et al., 2002]

According to them, digital records are "series of performances across time" [Heslop et al., 2002] —a *performance* being defined as "what is rendered to the screen or to any other output device", but also 'as 'a combination of a source and a process", or as "the collaboration of the source and process" [Heslop et al., 2002]. Digital records are performances because they

are “non stable artefacts” whose experience “last” only as long as “the technology and data interact” [Heslop et al., 2002].

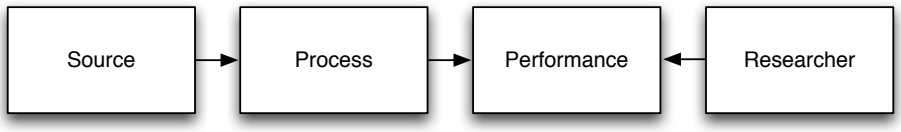

Figure 5.1: NAA Performance model – source and process components

This characterization is contrasted with the nature of paper records, “original and unique physical artefacts” that can be “experienced directly if [the researcher] can read the language” and “can only be experienced at one place in time”.

In order to address the relation between “equivalent performances” and “same records”, the concept of *essence* is introduced:

> “The project team developed the concept of a record’s ‘essence’ as a way of providing a formal mechanism for determining the characteristics that must be preserved for the record to maintain its meaning over time. The performance model demonstrates that digital records are not stable artefacts; instead they are a series of performances across time. Each performance is a combination of characteristics, some of which are incidental and some of which are essential to the meaning of the performance. The essential characteristics are what we call the ‘essence’ of a record.” [Heslop et al., 2002]

The NAA Performance Model introduces important yet challenging ideas that we can summarize as follows:

1. The known fact that digital resources are always mediated by technology here justifies the claim that “digital records are not stable artefacts” and that they *last* only when certain circumstances are met —i.e. for as long as “the technology and data interact” [Heslop et al., 2002].

2. When those circumstances are met, we should be able to recreate "equivalent performances" distinguishing between the essential characteristics of a record —those that qualify its *essence* and therefore, must be present in equivalent performances— and those that are contingent and not essential to the meaning of a record.

3. Each *equivalent performance* is a "view" of the *same record*, a "new original copy" that multiple people should be able to experience at the same time.

The NAA Performance Model explicitly addresses the need for something *concrete* that needs to experienced by a researcher in order access the *essence* of a record: a *performance*.

If performances are concrete objects, what kind of concrete objects are they?

From a technical perspective they might be described as a sequence of *states* in the machine, but this approach does not capture appropriately the user experiential perspective.

They appear to be objects spanning in place and time and whose identity conditions involve both material objects (such as an output device) and event–like components (consider, as an example, the performance of a digital movie); they are also objects potentially affected by the user input and interaction with the digital environment in place (consider the variation in different performances of a video game or other interactive media that are, in fact, essential to game play).

If we consider again a digital environment DE, any particular performance produced in DE at run time can be understood as a (potentially complex) situation part of DE itself that exists until the appropriate process lasts.

Consider a parallel example. For a person watching a movie, the movie *itself* cannot be reduced to the sum of its photograms: it is indeed perceived as an individual entity, a continuum. For similar reasons, performances cannot be reduced to the sum of components involved, being them physical objects and events, or to the sum of machine states. This account is also in line with the notion of performance as adopted in fields such as performing

arts and linguistics.[3]

Performances are situations that, while involving physical objects and events, are individuated as *whole individuals* bearing unity and identity conditions by a user experiencing and interacting with them.

When users engage with what we colloquially call *digital objects*, they do so by recognising, experiencing, and interacting with concrete performances (according to their sensory perception of the world).

The notions of *performance* and *essence* together seem to capture many of the characteristics that Thibodeau assigns to his *conceptual object*, distinguishing however between the concrete objects that users experience, namely *performances*, and the characteristics that objects of this sort need to share in order to be appropriate performances of the same record.

5.2.3 The InsPECT Project

A similar distinction and approach has driven the work of the The Investigating the Significant Properties of Electronic Content Over Time (InSPECT) Project[4] [Knight, 2009].

As pointed out by Heslop [Heslop et al., 2002], the interpretation of the same source within different digital environments can lead to the production of performances that are significantly different —even if potentially equivalent from the perspective of certain communities of users. InSPECT recognises this critical aspect when defining the notion of *significant*

[3]The notion of a performance is central and received proper attention and analysis in contexts different from digital preservation: for example, *performance* in the performing art sense and *performance* in the linguistics sense.

In the performing art sense,a *performance* is usually defined as "the act of performing" a musical play, a stage drama , a live installation, etc. On this account, performances seem to be concrete particular events in time.

An extensive ontological analysis of the concept of a performance has been carried out by Peggy Phelan and presented in her book "Unmarked: the politics of performance" [Phelan, 2012].

According to Phelan, a performance is a non–repeatable act —"Performance in strict ontological sense is nonreproductive [...] they live in the present" [Phelan, 2012].

There is another sense of performance that is potentially relevant to our context: in a linguistics sense, a performance is the "actual use of language in actual situations". A linguistic performance is broadly conceived as the production of an utterance of a sentence.

[4]See http://www.significantproperties.org.uk/.

properties:

> "[Significant properties are] the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record." [Knight, 2009]

This definition appeals to the following rationale:

> "An assumption implicit in the OAIS Reference Model is that a single type of Representation Information will exist for each Data Object that will be used to recreate the Information Object. Although ideal, this does not reflect practical experience of accessing a Data Object in a digital environment. As recognised by the National Archives of Australia in its Performance Model, it may be more accurate to recognise the existence of several Representation Information variants for a single Data Object. The use of one Representation Information variant may yield an Information Object that differs from that rendered by a second Representation Information variant. The differences between the two recreations may be considered minor or major dependent upon its influence upon the access, use and interpretation of the Information Object." [Knight, 2009]

The approach of InSPECT relativizes what successful preservation is: *significance* is defined with respect of the resource under consideration, but also with respect to the expectation and intent of stakeholders involved in the process. An approach very well captured by Dapper and Farquart's expression 'significance is in the eye of the stakeholders' [Dappert and Farquhar, 2009].

The notion of performance really places the assessment of successful digital preservation in the eyes (or ears, etc.) of users when they engage with technology. Similarly to bit–level preservation, we are facing a situation where an agentive role is central to preservation

assessment and there is no certainty of successful preservation if not *in the moment* when a performance is produced and a user (of a designated community) engages successfully with it recognizing the characteristics that constitute its *essence*.

In this sense, we are facing an account of preservation that resembles the *transitional* account of preservation presented above, where successful preservation is assessed according to an equivalence relation between concrete entities over time. The substantial difference resides, however, in the ephemeral nature of performances. Performances, while concrete, appear to be temporally–constrained, technology–dependant, and non–repeatable situation–like objects. This implies that any individual performance is unique and ceases to exist when specific processes at the machine level terminate: they do not *persist* in the same way material objects do.

'Preserving access to digital materials' appears then to be a metaphorical expression for 'producing equivalent performances over time'.

This approach, however, does not capture all the intuitions behind the way we talk about digital objects, nor certain assumptions in digital preservation.

5.3 What 'preserving access to digital materials' does mean

Let's consider first the notion of *digital material* itself.

5.3.1 What *digital materials* are, after all

Expressions like the '*same* e-book is available in E-Pub and Mobi', or again, 'a digital document is migrated from one file format to another' appear to assume the existence of something that is *the same* in different contexts, a thing that reflects Thibodeau's notion of *conceptual object*, but not quite. As observed above, there is an unfortunate abstract/concrete

conflation in the notion of *conceptual object*, partially solved by the notion of performance.

Terms like 'digital document' or 'e–book' or 'video game', however, do not refer to any particular performance, nor, strictly speaking, to "a series of performances across time". Performances are concrete non–repeatable entities whereas, intuitively, *the same* e-book can be read on different devices and *the same* video game can be played over and over again.

What sort of things are a *video game* or an *e–book*, essentially? What about objects that, intuitively, belong to similar categories (such as digital documents or digital pictures) that so often are identified as objects to be preserved?

There is an almost natural affinity between terms like 'video game', 'e–book', 'digital image' and the notion of 'digital material' often found in the digital preservation literature, one that suggests we cannot dismiss these intuitions completely.

We suggested above that *performances* are *concrete situation–like objects*: they are complex entities involving material objects, events, and relationships among them. Objects like video games, e-books etc. appear to be something like an abstract counterpart of their respective specific performances.

For these reasons, we propose here to characterize the notion of *digital material* in terms of *abstract state of affairs*[5] which can *obtain* in multiple concrete situations.

Generally speaking, a state of affairs is *a way things might be*, and state of affairs can be effectively understood as the abstract counterparts of *concrete situations*. We can also say that, when a state of affairs obtains through a situation, that situation makes the state of affairs *actual*.

Adopting Edward Zalta's approach [Zalta, 1983], we say that states of affairs *encode* properties and relationships, while concrete situations *exemplify* them.

A minimal state of affairs σ *encodes* an individual property ϕ. We can say that σ obtains if an only if there is a situation s such that $\phi(s)$. Which is also to say that a situation s

[5]See ¡http://plato.stanford.edu/entries/states-of-affairs/¿.

makes a state of affairs σ actual if and only if s (minimally) exemplifies every property ϕ that σ encodes.[6]

The notion of *states of affairs*, in this sense, closely relates to the notion of *proposition*: the sentence "the cat is on his mat" is not only said to express the proposition that *the cat is on his mat*, but also to *describe* the state of affairs of *the cat being on his mat*. In fact, there are similarities between states of affairs and propositions. Propositions are true or false; states of affairs obtain or not.[7]

On this account, a specific digital material can be understood as compound state of affairs encoding all the properties and relationships that a performance appropriate for that digital resource (i.e. a performance that makes the digital resource actual) must exemplify.

Consider a basic example about a digital textual document —i.e. a particular sort of *digital material*. Given a symbol structure tx (the text), and interpretive frame IF, a variable y ranging over temporal locations, and a variable x ranging over digital performances, a digital textual document can be characterized, minimally, as the state of affairs encoding the following relationship:

$realizes(x, tx, IF, y)$

To form the name of the *digital material* corresponding to the open formula, we bracket the formula with double solidi:

$//realizes(x, tx, IF, y))//$.

This digital material obtains if and only if a digital performance s is produced at time t_1 such that:

$realizes(s, tx, IF, t_1)$.

We believe that this notion of digital material as state of affairs captures precisely the

[6]We note here that logical compounds of state of affairs are themselves state of affairs. Any particular situation can make actual more that one atomic state of affairs, or the logical compositions of them.

[7]The fact that a state of affair can be described by a *sentence* plays a central role in how states of affairs can be modeled. For an account of the distinction between *being* and *being modeled as* see 2.7 in [Devlin, 1995].

intuitions behind the use of terms like 'e–book' and 'digital document'[8] without negatively affecting our practical intuitions about digital preservation —i.e. that preserving digital materials means producing equivalent performances over time. For example, a particular digital document encodes a series of characteristics that needs to be exemplified each time a performance of that digital document is produced in order to provide the user with the intended experience.

This approach is also in line with the idea that successful preservation can only be assessed with respect to the expectations of a designated community of stakeholders. The notion of *significance* [Giaretta et al., 2009, Dappert and Farquhar, 2009, Lynch, 2013] adopted in InSPECT attempts to capture these expectations in terms of characteristics that are significant in order for preservation to be considered successful. We can say that the production of a particular performance satisfies preservation requirements if and only if a produced performance makes actual an *intended* state of affairs —namely, the state of affairs that encodes all the individuals, properties, and relationships that have been identified as significant.

5.3.2 Preserving access to digital material: an account

Clearly, even on this account, traditional accounts of preservation do not apply: we are dealing again with objects of abstract sort that require no preservation efforts on our part to ensure their continued existence, or objects like performances that, while concrete, only exist at run-time and therefore are inherently ephemeral.

As recognized early on by the InterPARES community:

> "While the phrase "preserve an electronic record" is convenient and undoubtedly will continue to be used, in many variations it is a shorthand expression that

[8] It also applies, we believe, to objects like *digital pictures* or *video games* that manifest characteristics that cannot be reduced to propositional information, or representations of propositional information. The caveat here is to properly model within the framework the characteristics that performances of these kinds of material exemplify. The framework itself only models the relationship between performances and digital materials, not the specific characteristics that they, respectively, exemplify and encode.

> belies reality. Empirically, it is not possible to preserve an electronic record: it is only possible to preserve the ability to reproduce the record. That is because it is not possible to store an electronic record in the documentary form in which it is capable of serving as a record. There is inevitably a substantial difference between the digital representation of the record in storage and the form in which it is presented for use. It is always necessary to use some software to translate the stored digital bits into the documentary form of the record." [InterPARES, 2001]

While the focus of InterPARES are digital records, similar perspectives can be applied to other digital stewardship scenarios.

Producing equivalent performances over time means producing over time performances that make actual the same digital material, providing means for a user to access certain information.

Preserving the requirements for producing equivalent performances, however —in InterPARES terms "preserve the ability to reproduce the record"— is again a misleading metaphor: all the things involved in the production of a performance are either:

- objects of abstract kinds —for which preservation does not apply, nor is required— or
- material objects, like hardware components, that are in fact required to be replaced over time.

A different approach, and the one espoused here, may consider these issues from an information communication perspective.

Requirements to produce equivalent performances might be expressed in terms of requirements for certain information to emerge within a digital environment such that the machine agent can produce appropriate performances according to that information.

The general formalities of this process are not dissimilar to those that can be used to describe the requirements for any agent to be presented with information. In fact, a related perspective can also be applied to describe how information emerges for a human agent from specific performances, more precisely describing how "a data object [i.e. the bits] interpreted using its representation information yields information [for a user]".

5.4 Conclusion

We have seen in this chapter how the insights from different yet related conceptual approaches to digital preservation contribute to a more consistent theory of what digital materials are and what it means to preserving access to them.

The next chapter will introduce a perspective on information based on influential information and communication theories [Grice, 1957, Devlin, 1995]. The emerging framework will be applied to provide a digital stewardship paradigm that avoid the several misleading metaphors and category mistakes involved in the use of the term preservation.

CHAPTER 6

THE CONTINGENT AND INTENTIONAL NATURE OF COMMUNICATION

6.1 Introduction

Information, we argued in Chapter 4, is a thing that cannot be literally preserved and the problem of *preserving information* can be more effectively tackled adopting a strategy that confirms the intuition that preservation is about interpretation and access.

In Chapter 5 we discussed several influential attempts to develop this approach into a comprehensive account of digital preservation. While these contained deep insights that advanced our understanding, they also contained flaws that limited their achievement. The notion of *digital material* has been clarified and a provisional account of *preserving access to digital material* provided. The presented account involves, again, interpretation and access processes, ultimately suggesting that many issues in digital preservation could be more effectively described from a communication of information perspective.

This chapter elaborates on these topics in greater deal, addressing:

- the fundamental kinds of things that are involved in information representation and the contingent relationships among them;
- the essential role *agents* play in establishing these relationship when involved in informing scenarios where communication happens.

The emerging framework, informed by Keith Devlin's work in Situation Theory [Devlin, 1995, Devlin, 2006], can be applied to all contexts where the flows of certain information

content needs to be modeled, including those scenarios —such as those addressed in digital stewardship— where communication is mediated by digital technology.

6.2 Representing information

Keith Devlin notes:

> "The first thing to observe [about information] is that information is 'carried' or 'arises from' a representation. [...] Without some form of representation there can be no information. But just because the information cannot 'exist' without the representation, it does not follow that the representaion is all there is —that information is somehow *contained in* or *part of* the representation." [Devlin, 1995]

We follow here a similar approach: the assumption is that we can only access information in virtue of some form of representation of that information.

The notion of *representation* in the context of information, knowledge, and data representation, refers to the conventions for the arrangement of things in the world in such a way as to enable information to be encoded and later decoded by suitable agents and systems.[1]

By *levels of representation* we mean those levels of abstraction identified in linguistics, philosophy of language, and computer science that are useful to describe roles that concrete and abstract entities play in the process of information communication.

Modeling these levels of abstraction has also proven to be effective to describe information–carrying resources from an analytical perspective [IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998].

[1]This definition is adapted from here: http://guide.dhcuration.org/representation/.

We distinguish here between *concrete representations* and *symbolic representations*, contingent roles that might be played, respectively, by concrete and abstract objects[2] in the 'representation stack' of propositional information.

Previous work by the Data Conservancy Data Concepts (DCDC) group at the University of Illinois identified the kinds of things that might play these contingent roles and potential relationships among them.

6.2.1 The Basic Representation Model

The Basic Representation Model (BRM) [Wickett et al., 2012] developed by the Data Conservancy Data Concepts (DCDC) group at the University of Illinois provides a high level description of the *kinds of things* (and relationships among them) that play *representation roles* when propositional information is communicated.

As described in Chapter 2, BRM (Figure 6.1) consists of three entity types —*Propositional content*, *Symbol Structure*, *Patterned Matter & Energy*— and three relationship types —*Is Expressed By*, *Is Encoded By*, *Is Inscribed In*.

Important of BRM is a clear distinction between kinds (explicitly modeled) and roles, distinction that is essential for proper and consistent conceptual modeling.

Notions like *Propositional Content*, *Symbol Structure*, and *Patterned Matter & Energy* reflect established ontological kinds extensively investigated in philosophical ontology, providing solid foundational basics for their appropriate description and and their adoption as primitives in our discourse about digital preservation.

BRM, although recognising it, does not explicitly capture the contingent nature of these relationships, nor the causality that allow these relationships to be established in the first

[2]This distinction can be effectively aligned, respectively, with the notions of *analog representation* and *digital representation* described by Dretske in his book 'Knowledge and the Flow of Information' [Dretske, 1983] and reported by Devlin [Devlin, 1995]. In avoiding confusion on the use of the term 'digital' we adopt here a different terminology: the term 'concrete representation' will be used instead of *analog representation* and *symbolic representation* will be used instead of *digital representation*.

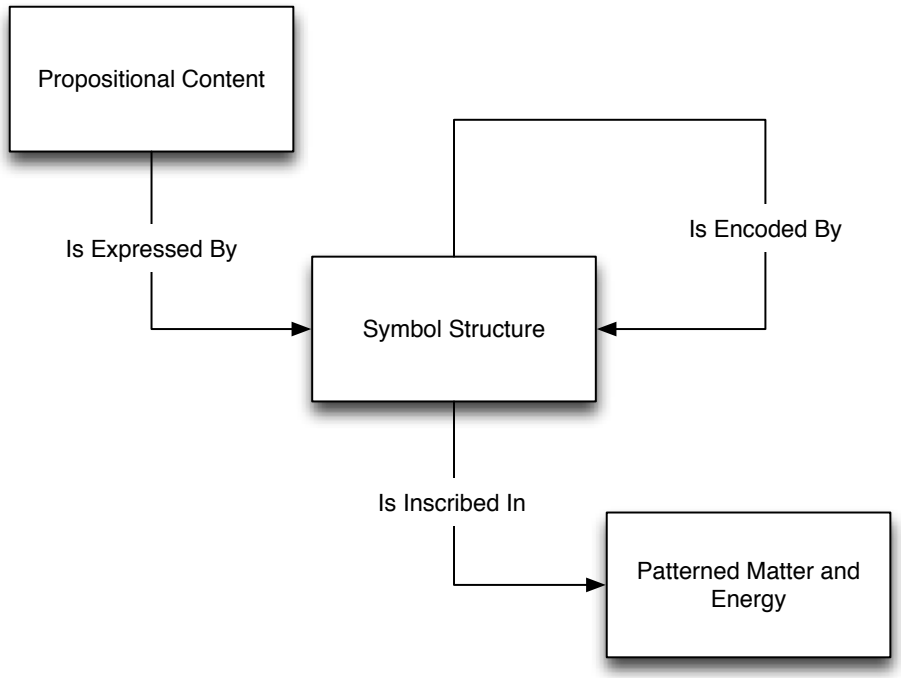

Figure 6.1: Basic Representation Model diagram

place:

> "[BRM] does not indicate how these objects came to enter into these relationships, spell out what entities and events are involved in creating and sustaining these relationships, or provide full details on how these events are situated in the context of [digitally–mediated] communication." [Wickett et al., 2012]

The Systematic Assertion Model (SAM), developed contextually with BRM, addresses some of these critical aspects for a particular kind of digital resource: digital data. SAM complements BRM by focusing on "key provenance events through which propositional content and symbol structures acquire the status of 'data content' and 'data', respectively." [Wickett et al., 2012].

Appealing to concepts derived from Situation Theory [Devlin, 1995, Devlin, 2006], in what follows we modify and extend BRM to capture the roles that agents play in scenarios where communication happens, scenarios where the relationship types identified in BRM are

instantiated and roles assigned.

6.2.2 The concrete representation level

Recall the examples of informing scenarios introduced in Chapter 4.

A Claudia engages a Chicago Transportation Authority (CTA) officer in a conversation to obtain information about buses to downtown. When Claudia asks the officer, the officer informs Claudia that *the bus 147 to downtown leaves at 4:15PM* (some information);

B Claudia picks at the bus station a CTA bus schedule printed brochure. By 'reading' the brochure, Claudia is informed that *the bus 147 to downtown leaves at 4:15PM* (some information);

C Claudia 'opens' a digital version (e.g. a digital brochure in PDF format) of the CTA bus schedule on her computer. By 'reading' the digital brochure on screen Claudia is informed that *the bus 147 to downtown leaves at 4:15PM* (some information);

Each of these examples accounts for different prototypical cases of informing scenarios: in example (A) Claudia is presented information in virtue of an intentional oral communication act by another agent; in example (B) Claudia is presented with information by engaging with an information–carrying physical artefact (a paper document); in example (C) Claudia is presented with information in virtue of her interaction with a digital performance.

In each of these scenarios, there are well individuated parts of the world Claudia attends to and interact with through her senses, and these parts of the world —a spoken utterance, a written document, and a performance of a digital material, respectively— function for Claudia as *concrete representations* of some information.

Situation Theory provides a primitive to characterize these (potentially complex) structured parts of the world: the notion of a *situation*.

The Situation Theory notion of a *situation* was first introduced by Barwise and Perry:

> "The world consists not just of objects, or of objects, properties and relations, but of objects having properties and standing in relations to one another. And there are parts of the world, clearly recognized (although not precisely individuated) in common sense and human language. These parts of the world are called situations. Events and episodes are situations in time, scenes are visually perceived situations, changes are sequences of situations, and facts are situations enriched (or polluted) by language." [Barwise and Perry, 1980]

While the notion of a situation has been refined in the Situation Theory literature in various, though related, ways, for the purpose of this work, we adopt the following definition:

Definition 3 (Situation) *A **situation** is a temporally and spatially located structured part of reality perceived as a unit by a cognitive agent in virtue of a scheme of individuation.*

On this account, a *scheme of individuation* is the way an agent "carves up reality" into cognitive "manageable pieces" from her "fuzzy perception of the world" [Devlin, 1995]. It is the (innate or acquired) cognitive capability of an agent to discriminate entities in the world as individuals —and particular situations are examples of such individuals.

Different classes of agent have different schemes of identification. Situation Theory is mostly concerned with agents of class *Men*[3] [Devlin, 1995] —i.e. the class of all human agents— but Situation Theory can be applied effectively to other types of agents, for example computing machines.[4]

[3]While agent of class *human being* can be understood as having a shared scheme of individuation, it is plausible to believe that, at a finer level of granularity, each different agent has some personal scheme that is not necessarily shared among the other agents in the same macro class.

[4]In [Devlin, 1995], Devlin illustrates an example of application on a simple robot.

From an agent's cognitive perspective, situations function as *concrete representations* from which information emerges —they *carry* information for an agent. More generally, however, situations are *contexts* we can attend to, talk about, and be immersed in —therefore affecting, contextually, what information is communicated.

The notion of a situation will be adopted later in this chapter instead of the notion of *Patterned Matter and Energy*. The reason behind this change is justified by the fact that in many circumstances —including those where communication of information in mediated by digital technology— what counts as a concrete representation cannot be reduced to a specific material object. In the case of a particular *spoken utterance* or a particular *digital performance*, for example, while certain *patterned matter* is involved in defining their identity and unity, for their inherent temporal aspect they cannot be reduced to any particular material object. The notion of a situation also captures the agent–and–context–dependent role of *being a concrete representation*, a role that certain structured part of reality plays: a situation (or at least its individuation as such) is functionally dependent on an agent "picking up" that situation from her perception of the world. In fact, even when a situation only involves a particular material object (as in the case of a paper document), the recognition of the object *as a document* (a situation of a certain sort) is agent dependent.[5]

A situation carries a particular propositional content for an agent in virtue of the agent recognition of certain abstract patterns from the situation, in such a way as to neglect details that cannot serve to differentiate meaning. These *abstract representation levels* will be discussed next.

[5]This account is not dissimilar to Suzanne Briet perspective of what counts as a document. As reported by Michael Bukland, Briet notes that "An antelope running wild on the plains of Africa should not be considered a document,[...]. But if it were to be captured, taken to a zoo and made an object of study, it has been made into a document. It has become physical evidence being used by those who study it". [Buckland, 1997]

6.2.3 The abstract representation levels

When they experience situations in the world, agents also recognize, through their *scheme of individuation*, recurring patterns (called *uniformities* in Situation Theory). These patterns allow agents to classify individuals within a situation (and situations themselves) with respect to similarities or differences. Situation theory adopts the notion of *type*[6] to characterize classes of individuals that share common characteristics.

In virtue of these recurring patterns, abstract objects are recognised. An example of this process is the capability of cognitive agents to recognise the *individual particulars* as instances of a specific kind (for example, a particular dog being an instance of the kind *dog*).

This process is described as the process of *abstraction* in Situation Theory.

Situation types in this sense can be effectively aligned with the notion of *abstract state of affairs* described in the previous chapter. Situation types abstract from particular concrete situations those characteristics that are functional to specific cognitive processes, one of which is the recognition that certain abstract objects stand in contingent relationship with situations. Examples of such objects are symbol structures, objects that function as representations of information at the abstract level.

Examples in this sense is the recognition of a sentence (an abstract symbol structure) from a spoken utterance (a situation), or a text (a symbol structure at a lower level than sentences) from a paper document (a situation as well). The process of abstraction also allows the recognition that a spoken utterance and a paper document can realize *the same sentence* —albeit at different level of encoding in the representation stack.

While the pattern recognition is enabled by the agent *scheme of individuation* (and happens at the perception level of cognition), the process of abstraction involves an interpretation of the sensory signal. This is actually a proper feature of cognitive agents: they are able to

[6] *Type* in this sense should not be confused with the metaphysical notion of *natural kind* nor with the formal ontology notion of *type*. A *type* in this sense simply denotes a class of individuals (e.g. a class of situations) sharing certain common characteristics.

go from the concrete to the abstract symbolic level such that discrete unit of information can emerge from the fuzzy perception of the world.

The outcome of this interpretation process is governed by specific *constraints* applied in the process.

Constraints are "natural laws, conventions, analytic rules, linguistic rules, empirical lawlike correspondences, or whatever" [Devlin, 1995] and, more generally, the background knowledge that an agent applies in the process of going from the concrete to the abstract level —ultimately leading to the propositional information level. Both the agent's *scheme of individuation* and the set of *constraints* an agent applies in the cognitive process are components of the agent's *interpretive frame* [Dubin et al., 2011].

Constraints typically operate over classes of situations sharing certain characteristics — i.e. at the level of *situation types*. And, in fact, constraints such as the English language vocabulary and grammar can be multiply applied to a variety of situations where certain patterns are recognized —being them patterns of sounds in spoken utterances or marks of ink on paper.

The process of abstraction also supports the recognition that a symbolic representation can be *encoding* other symbolic representations in a recursive set of contingent mappings. Any specific mapping —and the *encoding* relationships— are again governed by specific *constraints*. Linguistics and philosophy of language have extensively discussed and described such levels for both oral and written communication, an example being the *phoneme level* —that abstracts speech sounds in such a way as to neglect details that cannot serve to differentiate meaning. Other analogous kinds of abstractions (sometimes called 'emic units') include *morphemes*, *graphemes*, and *lexemes* [Pike, 1954].

What information emerges from a particular concrete situation is functional of the specific *constraints* the agent is "attuned to or at least aware of" [Devlin, 1995] and are applied in the interpretation process.

6.3 The agent's perspectives on information communication

When a cognitive agent is exposed to a situation, certain information sometimes emerges. What information emerges is a function of the *interpretive frame* —in terms of *scheme of individuation* and *constraints*— that the agent applies in the cognitive process.

This perspective aligns with the OAIS approach, where the requirements to access an *Information Object* are modeled according to a *Data Object* and its *Representation Information*: "*data* interpreted via its *representation information* yields *information*".[7] [CCSDS, 2002]

Following these premises, we can formalize the general requirements for an agent to be presented with a specific propositional content p —i.e. for an agent a to be successfully *informed* that p— and the conditions for successful communication between agents —i.e. the conditions for an agent a to successfully *inform* an agent b that p. Let's consider *being informed* first.

6.3.1 What it is to *be informed*

Given some propositional content p, an agent a is informed that p at time t_1 if and only if:[8]

(BI1) a individuates a situation s of type S at time t_1

(BI2) situations of type S function as a concrete representation of p according to IF

(BI3) a intentionally applies IF when experiencing s

[7] This initial view is expanded in the OAIS Reference Model, where other categories of information contribute to the meaning of data (for example *context information* and *provenance information* [CCSDS, 2002]).

[8] In ordinary circumstances, the interpretive frame is applied on the basis of certain assumptions and intentions that make it appropriately useful. In theory, there might be random applications of arbitrary interpretive frame that satisfy this definition; in fact, the exemplary cases are where an interpretive frame is chosen deliberately in virtue of certain assumptions and with the intention of a certain result.

From this account, several observations follow:

- not only an agent is required for information to emerge, but what information an agent is presented with is a function of the interpretive frame she applies to the process;
- agents applying different interpretive frames when experiencing the same situation (or situations of the same type) can be presented with very different information;
- an agent can be presented with the same information from different combinations of situations and interpretive frames.

This perspective carries direct implications for contexts where a situation is meant to function as a concrete representation for a particular propositional content. This is certainly the case when an intentional informing act is in place (such as with oral communication) but also in contexts where certain resources (such as those found in libraries, archives, and museums) are curated with the intent for them to function as designated carriers of certain information. Maintaining a resource available does not imply that the information a resource is intended to carry remains available.

It also captures the fact that the same resource can carry different (or more) information for different users. For example, an archivist versed in diplomatics (an example of background knowledge that can be applied as part of an interpretive frame) is more likely to obtain more information when exposed to an archival document than a person without that knowledge.

The diagram in Figure 6.2 modifies and extends BRM, summarizing the types of relationships that need to be instantiated at any point in time for an agent to be presented with some intended information (modeled as *propositional content*). The diagram should be considered a snapshot of the relevant binary relationships in play and it does nor reflect dependencies and combinations.

The relationship types *Experiences* and *Discriminates* are cognitive–established relationships between and agent and individuals in the world (both abstract and concrete) and are

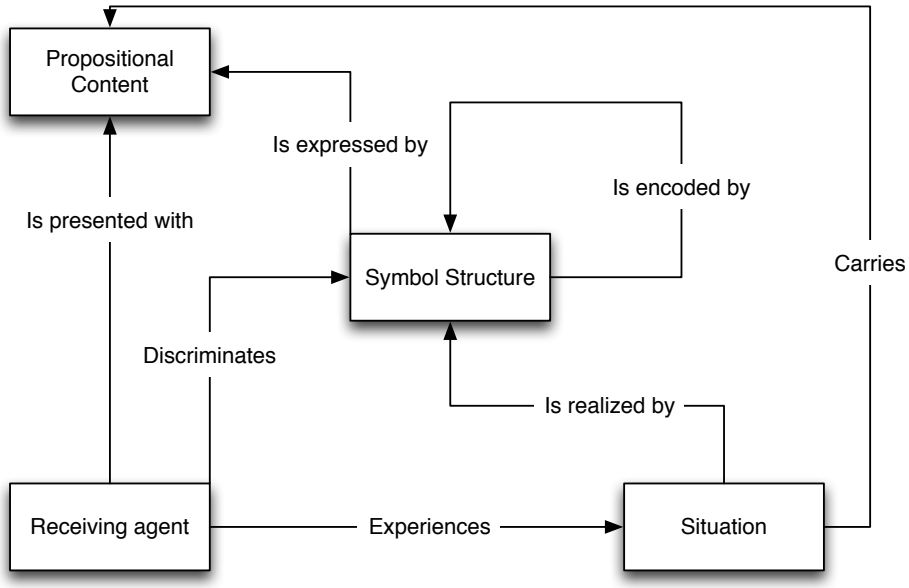

Figure 6.2: Entities and relationships

contingently dependent on the agent's *scheme of individuation*.

The relationship types *Is realized by*, *Is encoded by*, and *Is expressed by* and *Is carried by* represent, following BRM, the contingent mappings that are instantiated within the representation stack of information —relationships that allow to 'go' from a concrete entity an agent perceives through her sense perception up to the propositional content the agent is presented with. These relationships are contingent upon the agent involved in an informing situation and the *constraints* the agent applies in the cognitive process. They describe the way agents "carve up reality into manageable pieces" [Devlin, 1995] such that certain information emerges from the environment.

The relationship type *Is presented with* represents a general category of propositional attitudes between an agent and certain propositional content.

The diagram also accounts for the recursive *encoding* relationship type between symbol structures we identified in BRM. Such relationships capture the intuition that different *abstract levels of representation* can be individuated in the interpretation process, each level encoding another. This in not to say that each individually identified level might *directly*

represent certain information. Again the fact that certain propositional content is expressed by a symbol structure is dependant on the agent interpretive frame, therefore we need to acknowledge that for different agents (attuned to different interpretive frames) different information might emerge at any level.

The *Is carried by* relationship type is based on a function composition of the specific constraints mapping the individual representation levels and accounts for the relationship holding between a situation and certain information according to the entirety of a specific applied interpretive frame.

The distinction presented here reflects the contingent and intentional nature of these relationships:

- a *scheme of individuation* establishes contingent cognitive relationships between individual entities and the agent itself;
- *constraints* an agent is attuned are appealed to define how those individuals (both concrete and abstract) relate among each other in the informing process.

Let's consider now *what it is to inform* and the intentional participation of a *communicating agent* in the process.

6.3.2 What it is to *intend to inform*

The information an agent might be presented with does not necessarily arise from the experience of natural phenomena in the environment. In many situations of our everyday live there is an intentional communication process in place, one initiated by another agent who intends to communicate or *to inform*.

Here we explore what it means for an agent a to successfully inform an agent b that p, where p denotes some *propositional content*.

In identifying the conditions for 'a informs b that p' to be successful, we appeal to the Gricean theory of meaning [Grice, 1957], and the Situation–Theory–based paradigm presented above.

The Gricean analysis of speaker meaning runs as follows:

(G1) a utters x intending an agent b to form the belief that p

(G2) a intends that b recognises (G1)

(G3) a intends that b forms the belief that p at least partly because b recognises (G1)

We modify here the Gricean analysis to derive an account of *actual informing intent*.

In actuality, that 'an agent a intends to inform a potential agent b that p' is equal by definition to:

(II1) a utters x intending an agent b to be presented with p

(II2) a intends that b recognises (II1)

(II3) a intends that b is presented with p at least partly because b recognises (II1)

The changes introduced above are intended to 'loosen' some of Gricean requirements, in particular the requirements for the agent b to form a belief: the 'is presented with' expression represents a very general relationship between b and p avoiding unnecessary propositional attitude implications. They are also meant to emphasize an *actual* communication act, expression of an informing intent, regardless, for now, of a successful *informing process*, one requiring a receiving agent *being informed*.

Let's discuss each of these conditions in more detail.

Condition (II1): 'a utters x intending an agent b to be presented with p'

Condition (II1) describes the intentional act of an agent a to communicate some propositional content p.

In order for p to be communicated, the agent a must select (in the mind) a symbol structure x to function as a *symbolic representation* of p. A typical example in this sense is the use of a sentence in a language (e.g. the English sentence 'the snow is white') to *express* some propositional content (e.g. *that* the snow is white). Examples of concrete representations in this sense are spoken utterances but also, as we will see in the next chapter, persistent artefacts such as a written document.

The uttering event is understood as producing (or at least indicating) a situation s of type S intended to function as a *concrete representation* for p by realizing the selected symbol structure x. If a symbol structure x expresses some propositional content p for an agent a, and a situation s is the utterance of x, then we say that p is carried by s (for the agent a).

When an agent establishes these relationships, the agent appeals to an *interpretive frame* (or several *interpretive frames*) that govern the relational mapping between the entities involved. The established relationships are therefore *intentional* with respect to the uttering agent and *contingent* with respect to the applied *interpretive frame*.

On this account, situations of type S are those situations that make actual a state of affairs in the form

$$//\exists(y)[realizes(y, r, C_1^{IF}) \ \& \ expresses(r, p, C_2^{IF})]// \tag{6.1}$$

where the variable y ranges over concrete entities, r is a particular symbol structure, and C_1^{IF} is a constraint part of IF mapping a situation to a symbol structure x and C_2^{IF} is a constraint mapping the symbol structure x to the propositional content p. This state of affairs may obtain, or not, and it obtains only if the conjunction above is true.

Considering again the example of the English sentence 'the snow is white', what pragmatically happens is that the agent applies the constraints of the English language to establish relationships between an utterance situation and the sentence 'the snow is white', and be-

tween the sentence and the proposition that *the snow is white* (some propositional content).[9]

The diagram in Figure 6.3 summarizes the relationships between entities from an uttering agent perspective:

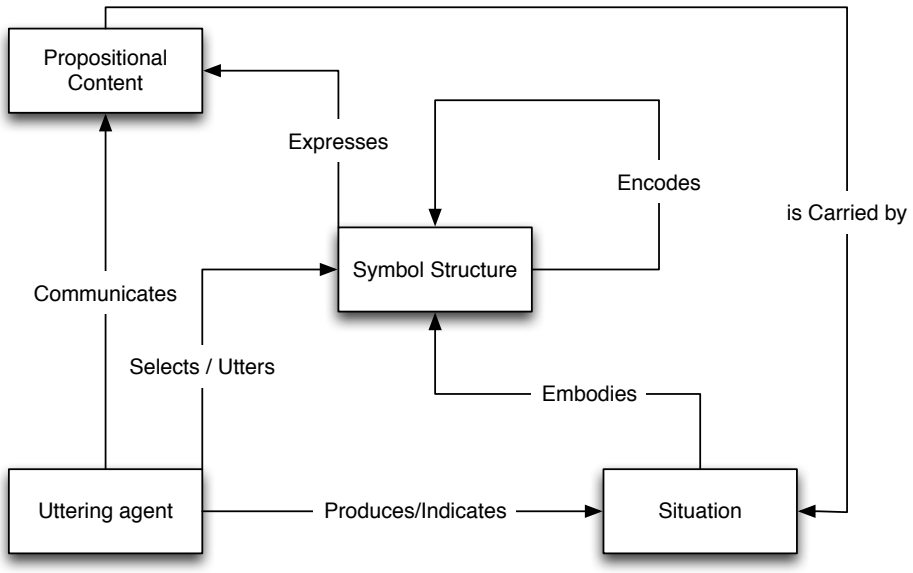

Figure 6.3: Entities and relationships

Similarly to the perspective on *being informed*, the relationship types *Is realized by*, *Is encoded by*, *Is expressed by* and *Is carried* express the contingent mappings that are instantiated in the representation stack of information according to specific *constraints* part of an *interpretive frame*.

Different from the perspective on *being informed*, the relationship types *Produces / Indicates*, *Selects / Utters*, and *Communicates* connecting the agent with the entity types describe the intentional process of 'building' the representation stack according to a particular interpretive frame.

[9] A similar perspective can be applied when multiple levels of abstract encoding working at different 'emic' levels are involved.

Condition (II2): a intends that b recognises (II1)

Conditions (II2) simply describes the intent of a communicating agent to let a potential receiving agent b recognizing her original communication intent. We interpret this condition here as a's intent to support (or at least not intentionally impede) b's recognition of the conditions under which a certain situation s is intended to carry p —i.e. the recognition by b of the *interpretive frame* actually applied by a in (II1).

Condition (II3): 'a intends that b is presented with p at least partly because b recognises (II1)'

Condition (II3) describes the causal conditions required for an *uttering act* to be an *informing act*. The agent a intends that a potential agent b is presented with p not by accident, but because of a *causal connection* between a's intent and b's relationship with p.

If b recognises a intention —i.e. b recognises (II1)— b becomes at least aware of the appropriate *interpretive frame* to apply when experiencing s —i.e. an interpretive frame sufficiently similar to one a applied in (II1). In this case, b is in a position to recognize that s carry p by vertically transversing the actual representation stack from the concrete representation level to the abstract levels of encoding, up to the propositional level.

From b's perspective, this process is perfectly aligned with the account of *being informed* presented above: an agent b is informed that p at time t_1 if and only if:

(BI1) b individuates a situation s of type S at time t_1

(BI2) situations of type S function as a concrete representation of p according to IF

(BI3) b intentionally applies IF when experiencing s

The additional criterium here is the recognition of an agent's *informing intent* that establishes *what* interpretive frame —in terms of *scheme of individuation* and appropriate *constraints*— should be applied.

We should finally note that the process of obtaining information from a situation is highly contextual and the interpretive frame actually applied in the process might be affected by a multiplicity of factors including, but not limited to, the context where the communication happens, other previous knowledge brought into bear on the interpretation process itself, etc.

This perspective is recognised by OAIS with its notion of *context information* [CCSDS, 2002], an essential component of any representation information. It is also recognised in Situation Theory and Situation Semantics where the notions of discourse situation —the situation where an utterance happens— and the notion of *resource situation* —a pre-existing situation mentioned, or referred to, in the utterance— are utilized as other parameters affecting what information is successfully communicated.

While these variances are recognised here, by interpretive frame we intend, when not otherwise noted, the *minimal* scheme of individuation and set of constraints required for certain information to emerge from a situation of a certain type.

6.3.3 What it is to *successfully inform*

From the perspectives on *being informed* and *(actually) intending to inform*, we can derive the conditions for a successful informing process.

For an agent a to successfully inform and agent b that p, the analysis runs as follows:

(SI1) a produces (or indicates) a situation s of type S by applying IF intending b to be presented with p

(SI2) a intends that b recognises (SI1)

(SI3) b is presented with p at least partly because b recognises that (SI1) and therefore applies IF when experiencing s

6.4 Conclusion

In this chapter we presented an analysis of information, describing the role that representation levels and agents play, and emphasising the intentional and contingent nature of information communication.

While information should always be distinguished from any abstract and concrete representation of that information, we must recognise that, in order to access and interact any information, an agent (such a user, but also a processing agent at the machine level) needs to engage with a concrete representation of that information.

An interpretation process is always required in order for an agent to be presented with certain information, a process that involves specific means —described in terms of *interpretive frames*— for the interpretation to happen as intended.

In the next chapter this perspective is applied to further analyze the issues of *information preservation*, concluding that many of the issues we face in the stewardship of our digital cultural heritages can be described from a communication point of view.

CHAPTER 7

INFORMATION PRESERVATION AS SUSTAINED RELIABLE COMMUNICATION

7.1 Introduction

In Chapter 3 we showed how definitions of *preservation* have evolved to address the changing nature of information–carrying objects, culminating in definitions of digital preservation where the emphasis was on maintaining access and not preserving objects.

In Chapter 4 we took a closer look at the common notions of *preserving the bits* and *preserving information*, noting how preservation in its literal sense does not apply, nor is required in the scenarios being described, and suggesting alternative accounts.

In Chapter 5 we explored three influential approaches to digital preservation that more accurately describe the complex nature of digital material. Despite the many insights, none of these approaches deliver complete and ontologically precise accounts. We also noted how the notion of *preserving access to digital materials* can be more precisely understood from a communication perspective: preserving access to digital materials is about sustaining interpretation and communication processes.

In Chapter 6 we elaborated on those topics by developing a set of interrelated concepts useful to understand how information is represented and communicated. The emerging conceptual framework builds on previous work by the Data Conservancy Data Concepts (DCDC) group at Illinois and is informed by Keith Devlin's Situation Theory ontology.

In this chapter, we utilize the framework as a lens to better understand the key issues and opportunities in digital stewardship.

Particular emphasis is given to the communication of information sustained by information–carrying artefacts and the roles of agents involved —especially in scenarios where changes in the underlying communication technology modify the requirements for particular information to emerge.

Core problems of identity and change in a digital context are addressed here, as are related problems of authenticity and integrity.

We concluded that successful digital preservation is best understood as

> *successfully sustained reliable communication through time and across changes in the mediating digital technology.*

7.2 Information–carrying artefacts: opportunities and challenges

As presented in the previous chapter, information communication is a highly interactive process where the parties engaged need to share some form of representation of that information and a sufficiently common interpretive frame.

Besides direct communication —such as face-to-face communication— communication often happens with the support of artefacts of some sort entrusted to function as information carriers. Examples from Chapter 4 describe variations on this kind of scenarios: Claudia is presented with information by engaging with a printed brochure (example B) and a digital brochure (example C) both functioning as carriers of some information.

Artefacts of this sort are commonly utilized to sustain the communication process across time and contexts, where no direct interaction between the artefact's producer and users is implied.

When an artefact is produced, or merely indicated, with the purpose of carrying some information, the intentional act assigns the artefact an *information carrying role.* This

process can be considered an *uttering act* similar in many respects to the production of a spoken utterance: the producer applies an interpretive frame such that the artefact counts as a situation (of a certain type) intended to carry information.

Substantially different, however, is the output of the process: the produced representation (and thus the informing situation) is not ephemeral, as a spoken utterance is: artefacts of this sort typically manifest some form of *persistence*. At future times then, agents can be presented with the information the artifact was intended to carry (at the time of its production or indication) if attuned, or at least aware of, an appropriate interpretive frame.

This form of communication is inherently open–ended, resembling a broadcast: if an artefact is produced by an agent a at time t_1 with the intention to inform that p, the informing process only resolves when an agent b (at a time t_2, future to t_1) engages with the artefact and is presented with p.

When information–carrying artefacts are involved in a communication process, new opportunities emerge: the very fact that material traces of information can be utilized to carry information into the future is at the very core of many forms of 'information preservation':

> "Humans have a long tradition of retaining information artefacts for future use. Collecting institutions —libraries, archives and museums— manage extensive collections of materials that can communicate events, insights, facts and perspectives to those who encounter them." [Lee, 2012]

However, new challenges are introduced as well. Consider again the Gricean–based account of successful information presented in the previous chapter, and in particular conditions $SI2$ and $IS3$.

(SI1) a produces (or indicates) a situation s of type S by applying IF intending b to be presented with p

(SI2) a intends that b recognizes that (SI1)

(SI3) b is presented with p at least partly because b recognizes that (SI1) and therefore applies IF when experiencing s

Conditions (SI2) and (SI3) require the recognition of (SI1). However, we cannot assume that an artefact in itself can support such a recognition.

Information artefacts can be produced according to interpretive frames that are common enough to be considered part of a shared background knowledge of the community where the artefact will be utilized. Intuitive examples are written documents, such as newspaper articles or books, where an adopted language[1] (e.g. the English language) is a shared enough interpretive frame such that the author's communication intent is recognized.[2]

Many times, however, such a recognition cannot be taken for granted. Christopher Lee summarized this perspective when discussing the role of institutions with respect to the 'curation of information':

> "The curation of information artifacts is fundamentally different from direct communication in that one cannot assume that the parties who generated the traces will be available to provide 'answers to further questions.' Instead, one must make educated guesses about likely uses of the traces and then pre–emptively respond to likely questions by embedding appropriate information in the mechanisms used to manage and provide access to the traces (e.g. repositories, collection descriptions, information packages)." [Lee, 2012]

Lee captures the central role that stewardship institutions need to play in sustaining the flow of information from the creator of an information–carrying artefact to its potential

[1] And other components of an interpretive frame such character mappings and layout conventions, among others.

[2] Even where a shared language is adopted, *contextual information* [CCSDS, 2002] might be required for the intended information to emerge. Consider for example a technical manual lacking a glossary of the adopted technical terms. While a reader might assume a commons sense meaning for those terms, technical terms are often laden with implicit and context-sensitive semantics that cannot be directly inferred from the document itself.

future users: beside 'preserving' an artefact,[3] they need to support and facilitate condition (SI2) such that (SI3) can successfully obtain without a direct interaction between the parties involved in the artefact–supported and open–ended informing process.

This perspective certainly applies to artefacts intended to sustain *directly* a human–level communication when they are experienced. Examples of these persistent artefacts are written documents, such as manuscripts, printed books, newspaper articles, etc. These sorts of persistent material artefacts are produced according to constraints that map level of representations human beings can typically discriminate and recognize.

A similar analysis can be applied to information communicated with the support of digital technology, albeit accounting for additional levels of complexity and the active mediation role that machine agents play in the process.

Let's explore both these scenarios in more details.

7.3 Communication sustained by *persistent material artefacts*

Consider the case of an original manuscript document, and how such a material artefact can sustain information communication.

When the document's author, we call her agent a, inscribes specific ink marks on paper with the intent of communicating some information p (an uttering act), she brings into being an utterance situation s of type S involving the produced artefact.

An interpretive frame, we call it IF, is required and applied in the process. IF includes a's *scheme of individuation* and multiple *constraints* mapping the different levels of encoding in the representation stack. On this account, S is the type of situation sharing certain characteristics such that situations of type S (e.g. s) are capable of carrying p according to

[3]Preservation here is intended in a broader sense as in the general definitions of preservation presented in Chapter 3.

IF.

When an agent b attuned to IF experiences s, s is recognized as being of type S and, if the appropriate constraints that are part of IF are applied in an interpretation process, b is presented with p.

So far we assumed that b is informed that p by engaging with *exactly the same* s the uttering agent a produced.

Because interpretive frames operate at the situation type level, however, we shall note that any situation x can be utilized in the process as long as x is of type S.

Different material objects can in fact exhibit the minimal set of characteristics such that p can emerge when sufficiently similar IFs are applied —examples being copies of an original manuscript produced by medieval monastic scribes, but also different copies of the same edition of a printed book. We adopt the following notation from Situation Theory to denote a situation x of type S:

$$x : S \tag{7.1}$$

Which characteristics are essential for any x to provide access to p in virtue of an appropriate interpretive frame IF is, in fact, dependent on the IF under consideration. The total constraint component of any IF mapping p to a situation s can be understood as a *function composition* of the individual constrains mapping the relevant level of abstraction in the representation stack.

When individually considered, as in the example presented in Figure 7.1, different *types* can be individuated selecting relevant parts of a representation stack starting from the propositional level, and according to the appropriate constraints mapping the levels.

Each type S_1, S_2, and S_3, respectively, imposes additional characteristics on a situation in order for that situation to be considered of that type: S_1 only requires that there be a

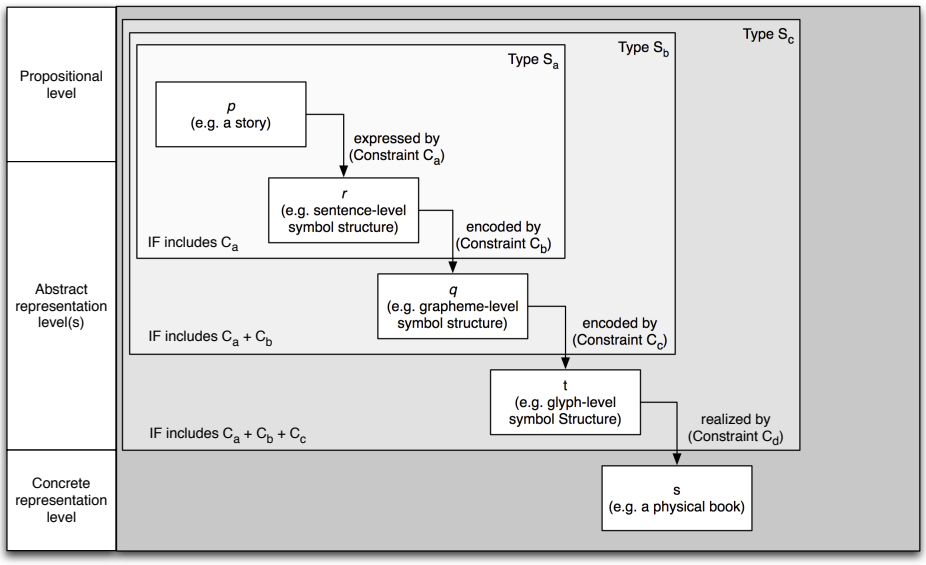

Figure 7.1: Types according to levels of representations and constraints

particular sentence–level representation (leaving open how that sentence is further encoded at lower symbol structure level, down to the concrete representation level); S_2 further imposes a certain grapheme–level representation (but not how graphemes are encoded in terms of glyphs); and so on.

Representations at the concrete level can in fact be *substantially different* still being considered of the *same relevant type*.

This analysis is here applied to explain:

1. preservation scenarios, such as those described in the *transitional definitions* of preservation presented in Chapter 3, where media migration and reformatting is applied;

2. more generally, how information can *intentionally* flow across a communication lifecycle where that information is represented differently at different times.

We shall note, before moving forward, that *transformative preservation practices* —those that somehow change the representation stack for certain information— are always expression of a *preservation intent* where established procedures are in place to ensure a certain outcome — i.e. information is 'carried over' to new concrete representation not by chance.

Given the inherent dependency of the success of those practices on specific interpretive frames, the role of agents in the process is obviously essential.

7.3.1 Preserving information: media migration and reformatting

Consider, for example, the practice of media conversion or reformatting where materials are preserved "in a condition suitable for use,[...] in a form more durable" [Reitz, 2004b] or "in a condition suitable for use [...] in a more persistent format, while leaving intact the objects intellectual form." A *microfilm* produced from a *printed book*, for example, under most circumstances is considered capable to carry the same relevant information content of the original book, while being substantially different from the printed book in the set of characteristics it exhibits. Media reformatting can be understood as altering part the representation stack of certain information, leaving intact relevant abstract levels and the relationships between them according the specific constraints.

Interestingly, successful media reformatting does not entail that an agent responsible for the operation is presented, in the process, with the carried information: these changes in the representation stack are typically applied at a 'lower' level of the stack where only symbol structures are involved. Different from the process of *translating* certain propositional content from, for example, one natural language to another, this is more a mechanical mapping that does not necessary involve high level cognitive processes —such as those required to operate at the propositional level where a change is introduced in the mapping between a particular *propositional content* and an expressing *symbol structure*.

7.3.2 Preserving information through a communication lifecycle

The observations presented above also allow us to consider how certain information can flow through a communication lifecycle, where a sequence of communication contexts —involving different agents, activities, and means to represent information— are interconnected such

that information systematically flows from one context to another.

Let's consider a lifecycle example: the production of a printed book from a publishing process perspective. A specific copy of a printed book (a copy you can find in your library) is the final outcome of a complex production process where the content from an author's original manuscripts flows across different contexts.

Let p be the propositional content the author intends to communicate —the one carried by the original manuscript. Consider now a sequence of concrete representations —the ones involved in the production process— and their respective types:

$$\{s_1 : S_1, s_2 : S_2, s_3 : S_3, \ldots, s_n : S_n\} \tag{7.2}$$

Consider also the sequence of appropriate interpretive frames to let p emerge from those representations, according to their types:

$$\{IF_{S_1}, IF_{S_2}, IF_{S_3}, \ldots, IF_{S_n}\} \tag{7.3}$$

Finally, consider a sequence of agents, each of them involved at a different step in the production process:

$$\{a_1, a_2, a_3, \ldots, a_n\} \tag{7.4}$$

The production process can be described as a sequence of contexts where agents mediate the flow of information between each other with the support of persistent representations of certain types. Figure 7.2 shows a simplified representation of ideal steps in the process, where different agents are involved in mediating the flow of certain information across different contexts where information is represented differently. In this simplified perspective, each agent engages with two different representations (possibly of different types), the first one

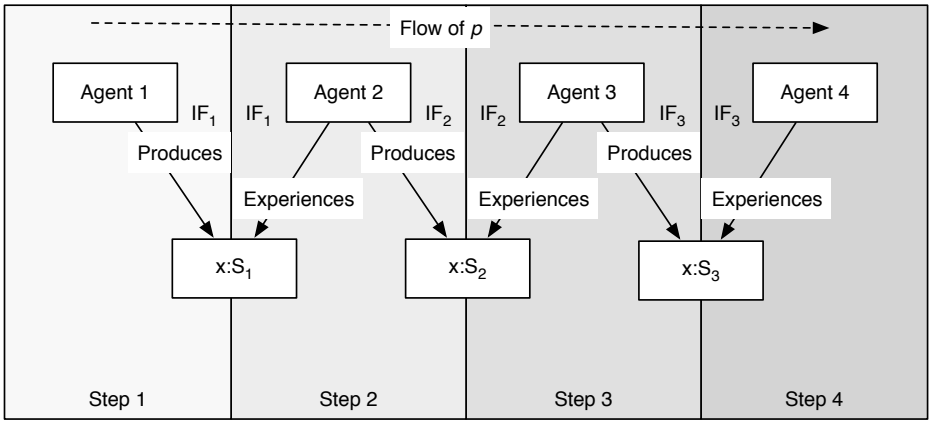

Figure 7.2: The flow of information p across different informing contexts

'carrying over' information p from a previous step —from which the agent is presented with p— the second one produced by the agent to 'carry over' p to the next step in the production process.

In order to sustain the flow of p from one context to another, agents involved at any step in the process need to be attuned to at least two interpretive frames. Consider for example the agent a_2 involved in step 2. The agent a_2 needs to be attuned, and to apply to the process, two different interpretive frames:

1. IF_1 to obtain p from $x : S_1$

2. IF_2 to produce $y : S_2$ such that $y : S_2$ carries p

Agents play a fundamental role in the process of sustaining the flow of information across different contexts where information is represented differently: p can successfully flow across contexts with different representation stacks involved only in virtue of agents capable of 'bridging' those contexts by being attuned to distinct appropriate interpretive frames.

The flows itself relies on the mutual recognition between agents of the interpretive frames that need to be applied in the process following the Gricean–based account of successfully informing, where conditions (SI2) and (SI3) need to be satisfied.

We shall finally note that, similarly to the media reformatting, the flow of information across multiple representations can sometimes be sustained through mechanical mapping between lower levels of symbol structures, where the semantic propositional level is not involved. This depends on the *types* of situations involved in the process.

7.4 Communication sustained by *digital artefacts*

When information is communicated with the support of digital technology, information is carried, at the user level, by what we called *digital performances*. As noted in Chapter 5, performances are *concrete*, yet *ephemeral*, entities and their production and existence are dependent on appropriate processes happening within a digital environment. Performances often involve temporal, dynamic, and interactive aspects, making them cognitively complex concrete objects that only the notion of a *situation* seems to capture appropriately.

Performances in this sense are not *persistent*: they are produced when an appropriate process happens at run time and cease to exist when the process stops. In this sense, when appropriately produced, each performance of a digital material is a different performance [Heslop et al., 2002] (of the same digital resource).

Despite their lack of persistence, performances function as *concrete representations* for certain information intended to be communicated to the users engaging with them.

From a preservation perspective, we cannot materially preserve a performance the same way we would preserve a material artefact: its parts are inherently temporal and transient. This quote from InterPARES is helpful again in describing the problem:

> "Empirically, it is not possible to preserve an electronic record: it is only possible to preserve the ability to reproduce the record. That is because it is not possible to store an electronic record in the documentary form in which it is capable of serving as a record. There is inevitably a substantial difference between the

> digital representation of the record in storage and the form in which it is presented for use. It is always necessary to use some software to translate the stored digital bits into the documentary form of the record." [InterPARES, 2001]

Described here is one of the peculiar characteristics of the communication mediated by digital technology: what counts as a *concrete representation* at the user level is, in fact, not a persistent artifact: it is not something that can be "stored" beyond any individual run-time session where a performance is produced.

While we are making advances by recognizing the distorting effects of metaphors and idioms, we need to be careful not to place too much weight on new distorting metaphors — however much insight they may provide in context, they still do not provide a reliable mode. Just as no entity is preserved, no ability is literally preserved either, and 'reproducing a record' is not, literally, producing a record, let alone reproducing one, but rather producing a performance that exemplifies the *documentary form* [Duranti, 2005] of a record. This process happens within a digital environment and involves a machine agent being presented with information other than the one a user is intended to be presented with.

This information —typically involving some form of machine level instructions— is carried by what InterPARES calls "digital representation of the record in storage" [InterPARES, 2001]. More precisely, this information is carried by a persistent representation —typically some form of *storage memory*— and expressed (directly or indirectly) by sequences of bits realized by the storage memory.

An interpretation process needs to happen according to a specific *machine–level interpretive frame* that allows a machine agent to map a storage situation (the concrete representation level) to the bits and other higher–level data structures (the abstract representation levels) and ultimately to the instructions (the information level). The machine agent acts upon this information to produce a performance.

This interpretive frame is typically expressed as *technical specifications* at various levels

including, but not limited to, *filesystem specification* and *file format specification*, each of them, respectively, constraining the mapping between a *storage situation* and a particular *set of bits*, and the mapping from the *bits* to the *instruction information*. The machine agent can become attuned to such an interpretive frame in virtue of appropriate software components within the digital environment that becomes available at run–time.

Even in this case, and in alignment with the perspective on communication presented in the previous chapter, constraints operate at the level of *situation types*. Consider a type ST_{bs} identified by the following state of affairs:

$$//realizes(x, bs, C_1^{IF_m})// \tag{7.5}$$

where bs is a particular bit sequence, x is a variable ranging over storage situations, $C_1^{IF_m}$ is a constraint (part of a broader interpretive frame IF_m available at the machine level) in virtue of which bs is realized by a storage situation x.

Every storage situation x that satisfies the formula '$realizes(x, bs, C_1^{IF_m})$' is a storage situation of type ST_{bs}; these are the storage situations that minimally exemplify a set of characteristics such that bs can be discriminated by a machine agent attuned to C_1^{IF}. Again, to form the name of the state of affairs corresponding to the open formula we bracket the formula with double solidi.

We shall stress here that different concrete storage situations of type ST_{bs} can realize bs for a machine agent attuned to IF_m (or at least attuned to the constraint $C_1^{IF_m}$ part of IF_m). This supports the intuition discussed in Chapter 4 that *the same* bit sequence, being an abstraction, can be multiply and repeatedly realized by different storage devices.

When the appropriate IF_m is applied in all its components, p_m emerges for the machine agent from a storage situation of type ST_{bs}. By acting upon p_m the machine agent produces a signal that changes the physical arrangement of an output devise, producing a performance,

we call it dp_1, that would make actual an intended digital material dm.

If we call DP_{dm} the type of those performances that would make actual dm, we can say that the interpretive frame IF_m links the storage situation type ST_{bs} to a particular digital performance type DP_{dm}.

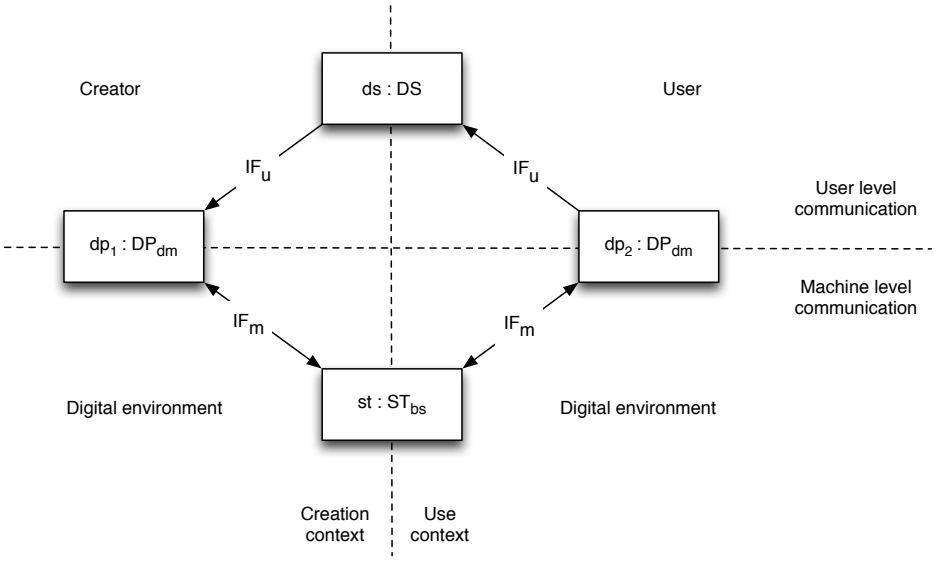

Figure 7.3: Situations at the user and machine communication levels

What a particular IF consists of is a matter of decisions taken at the time of 'creation' of a particular digital material[4] where an agent selects:

1. how to represent certain information via performances of a certain digital material —i.e. situations that makes actual an intended state of affairs— by selecting and applying an interpretive frame;

2. how the information to produce appropriate performance of the intended digital material is encoded and stored at the machine level (i.e. which sequence of bits needs to be realized by a persistent storage situation) and what machine–level interpretive frame (i.e. what file format specification) shall be applied at future times to a storage situation of that type to obtain the intended machine–level information.

[4]Or subsequent changes in the representation stack that might have happened later in the communication lifecycle of certain information.

From this perspective we can say that, given an intended digital material dm, DP_{dm} is the type of those performances that make dm actual.

The relationship between a specific bit level representation and a digital material can then be expressed as a relationship between types of situations governed by an appropriate interpretive frame, where situations of a certain type —e.g. storage situations of type ST_{bs}— carry information about situations of another type —e.g. performances of type DP_{dm}— when an appropriate machine–level interpretive frame —e.g. IF_m— is applied.

Similarly, we can say that performances of a certain type DP_{dm} carry information about a described situation DS when an appropriate user–level interpretive frame is applied.

The diagram in Figure 7.3 summarizes this perspective presenting both the *creation* and the *use* context.

This analytical approach can be applied to model prototypical challenges in digital stewardship.

7.4.1 Bits preservation

As described in Chapter 4, *preserving the bits* can be understood as preserving a certain analog arrangement of physical material[5] from which a computer can derive a signal of some sort that by relevant *conventions* is given a *digital interpretation.*

On this account, *preserving the bits* involves an active process where a machine agent within a digital environment discriminates a symbol structure (i.e. the bits) from a storage situation in virtue of the application of an appropriate interpretive frame.

In the long term, however, hardware components at the machine level are prone to change. Therefore, from a long term stewardship perspective, we should account for a change in the storage situations intended to realize the bits: bits need to be 'copied' from one storage

[5]No assumptions here are made on whether the physical material arranged in such a way is the exact same material at different times. The emphasis here is on the required availability of such an arrangement in order to access the bits.

device to another in a physical medium migration process.

Physical media migration — i.e. 'copying the bits'

The relationships between representation levels in the representation stack of certain information are contingent and dependent on specific interpretive frames that an agent applies when experiencing a concrete representation in the form of a situation.

As also noted, information can be intentionally 'carried over' across different concrete representations —involving different abstract representation levels as well— in virtue of agents attuned to appropriate interpretive frames. When an agent is presented with certain information from one concrete representation (in virtue of the application of certain interpretive frames), the agent can apply a different interpretive frame to produce or indicate a different concrete representations intended to carry the same information.

This perspective is intuitively applied in our everyday life, when, for example, we 'copy' a text from a physical book to our notepad with the intent for the 'copied text' to express the same information carried by the text realized by the physical book.

This perspective, however, can also explain the process of 'copying the bits' from one physical medium —e.g. a storage device– to another. This process can only happen within a digital environment where a machine agent can discriminate the intended bit sequence from a storage situation according to an appropriate interpretive frame, and arranges a different storage situation in such a way that, according to a (potentially different) interpretive frame, the same bit sequence can be discriminated.

Even at this low 'physical' level, we shall stress the machine agent's interpretation role in the process: bits can be 'copied' from one storage device to another only in virtue of a machine agent that is attuned to and applies potentially different, yet appropriate, interpretive frames. The process itself is part of on active communication process happening at the machine level, one that the machine agent sustains and mediates.

7.4.2 Digital materials preservation

Bit preservation is only the first required step for successful digital stewardship. Interpreting the bits such that an intended digital material obtains through appropriate performances is essential as well.

Also, as digital technology changes, or to accommodate specific *preservation intents*, digital materials might need to be *migrated* to new file *formats*.

The process of *preserving digital materials* is in fact a mediation process intended to sustain certain communication at the machine level in a predictable and reliable way and across time and changes in digital technology.

Let's consider the case where no change is expected nor pursued at the bit representation level, and then the case where digital materials are migrated to different file formats.

Producing appropriate performances over time from the bits

As we saw above, the process of producing a performance that appropriately makes an intended digital material actual can be effectively understood as a communication process where a machine agent is presented with certain instruction information by interpreting the bits according to appropriate constraints. These constraints, typically expressed in terms of *file format specification*, can only be applied when software components of an appropriate class are recognised by the machine agent (another discrimination process) and utilized in the interpretation process. The outcome of such a process is typically a signal intended to physically rearrange an output device (e.g. a screen, a set of speakers, etc.).

Despite the application of the same interpretive frame to the same bit level representation, performances produced over time can be significantly different, depending on the specific digital environment involved. Nevertheless, when certain minimal characteristics in the digital environment are met, performances produced in this way should be appropriate with respect to an intended digital material: they should minimally share (i.e. they all should

exemplify) the characteristics that the intended digital material *encodes*.

From a digital stewardship perspective it is therefore essential to capture and record (for example using appropriate *technical metadata*) the minimal requirements to produce appropriate digital performances given a bit level representation, both in terms of interpretive frames, but also in terms of a set of hardware and software components that are appropriate for the application of that interpretive frame.

Requirements in this sense should sustain the *flow of instruction information* that needs to happen at the machine level, and in such a way that the communication process can be multiply repeated over time, and accounting for variations in the digital environment involved in the process.

Format migration as mediated communication

Under certain circumstances, however, a specific *digital material* needs to be *migrated* from a file format to another. Examples of these circumstances include, but are not limited to:

1. the original file format is not optimal to support the production of appropriate performances in the long term (e.g. the original file format is a *proprietary format* associated only with a specific software application not maintained anymore);

2. the effective access to a specific digital material from members of a designated community —or the access to the information the material is intended to carry— depends on a hardware/software environment other than the one the original file format supports (e.g. users of a community are expected to be served digital images on the web, but the original file format of the images is not supported by modern browsers; or, scientific data encoded in a specific file format can be more effectively analyzed when the same data in encoded in a different format).

From a high–level perspective, the process of *migrating* a digital material from one file format to another is a communication process where a machine agent modifies certain mappings in the machine–level representation stack of certain information, such that information can flow across contexts where it is represented differently.

This process can be modeled, again, in terms of a N:1 relationship between different storage situation types (defined in terms of different bit–level representations) and the same intended performance type (defined in terms of essential characteristics that performances need to exemplify in order to make actual an intended digital material).

Given a type of storage situation ST_{bs_1} —the type of those storage situations that realize the bit sequence bs_1— and the type of digital performance DP_{dm} —the type of those digital performances that make actual a digital material dm— we said that a link between these type is established according to interpretive frames of a particular type IF_1^m —those interpretive frames that would let the appropriate instructions emerge from the bits such that a performance of type DP_{dm} is produced.

Appropriate software applications not only support the interpretation process to produce a performance of type DP_{dm} when an interpretive frame of type IF_1^m is applied to a storage situation of type ST_{bs_1}. They also support the process of encoding differently the instruction information necessary to produce performances of the same type DP_{dm} —therefore *producing* a storage situation of a different type ST_{bs_2} that realizes a bit sequence bs_2 that is linked to DP_{dm} according to interpretive frames of a different type IF_2^m.

An example of such format migration process is the migration of a textual document from the Microsoft Office Open XML format to the Portable Document Format (PDF) format, where many experiential characteristics are retained (if not the same interactive functionalities).

This process is not dissimilar in its logical components to the one presented above on the production of a book —where we described how human agents can sustain the flow of

information across different contexts where information is represented differently: the file format migration process can only happen when a (machine) agent is involved and it *mediates*, according to established procedures, the required communication process. Similarly as well, the agent needs to be attuned to, and apply, two different interpretive frames:

- an interpretive frame of type IF_1^m to obtain the instruction information from a storage situation of type ST_{bs_1};
- an interpretive frame of type IF_2^m to produce a storage situation of type ST_{bs_2}.

Again, it is important to stress that what set of characteristics a particular digital material dm encodes —and therefore appropriate performances of type DP_{dm} need to minimally exemplify— should be understood contextually, with respect to:

1. preservation *expectations* of the members of a community against which successful preservation is assessed —that might inform and affect stewardship decisions;
2. the *preservation intent* of a stewardship organization;
3. the social and technical constraints the organization operates within —in terms of policies, resources, and capabilities .

In fact, when considering performances produced by the interpretation of a storage situation of type ST_{bs_1} according to an interpretive frame of type IF_1^m, the type DP_{dm} —and therefore the characteristics that performances of that type need to minimally exemplify— can be defined differently in different contexts according to the criteria presented above.

On this account, frameworks such the one developed within the InSPECT project [Knight, 2009] to identify characteristics that are *significant* [Hedstrom and Lee, 2002, Dappert and Farquhar, 2009, Lynch, 2013, Sacchi and McDonough, 2012, Wickett et al., 2012] to assess successful preservation can play a major role in the process.

From an *information preservation* perspective, the conceptual framework developed here can be utilized as an analytical tool to individuate *specific levels in the representation stack of certain information* that need to be retained across format migrations such that information preservation can be considered successful with respect to preservation expectations of different communities and different stewardship intents.

The process involves analyzing the required human–level *interpretive frame* necessary to let the intended information to emerge from performances of a certain type, and decomposing the interpretive frame in terms of the particular *constraints* mapping the individual levels of abstraction —a process in many respects similar to the one described above for material artefacts.

We shall also note that, when transformative preservation actions —such as format migration— are applied, *provenance information* shall be captured as well utilizing appropriate metadata standards.

Also, to fully support an appropriate interpretation of performances at the user level — i.e. the application of an appropriate interpretive frame— additional *contextual information* [CCSDS, 2002,Lee, 2011,Lee, 2012,McDonough, 2011,McDonough, 2012] might be required, and shall be captured as well.

7.5 Conclusion

In this chapter we applied the conceptual framework of information representation and communication developed in Chapter 6 to analyze problems in information preservation. We started with information carried by material artefact intended to function directly as information–carriers, and we showed how many prototypical challenges and opportunities in digital stewardship can be made vivid and explained through the same analytical lens.

Successful digital preservation of information can, in this sense, be conceived as *sustained*

and reliable communication mediated by digital technology and agents involved in the communication process, where 'successful' is always contextually defined with respect to certain user–level expectations and stewardship intents.

The perspective presented here consolidates, within a consistent and precise conceptual framework, many of the insights from previous conceptual modeling approaches to digital preservation developed within the digital stewardship community [Thibodeau, 2002, CCSDS, 2002, Heslop et al., 2002, Duranti, 2005, Knight, 2009].

CHAPTER 8

CONCLUDING REMARKS

The notion of *preserving digital information* is fundamental in the broad digital and data stewardship agendas [Task Force on Archiving of Digital Information et al., 1996, Duerr et al., 2004].

In this research we showed how the notion of *preservation* itself, derived from the traditional library, archive, and museum practice of "minimizing deterioration or damage" of physical artefacts, is a metaphor that can mislead us to 'what is really going on' when the goal is to provide access to the same information over time.

We also showed how influential conceptual approaches developed within the digital stewardship community attempted to remedy some of these issues by analyzing the nature of 'digital objects' and their relationship to the information they are intended to carry. Despite the many insights, none of these approaches is complete and consistent: they still rely on loosely defined terms and underdeveloped conceptual foundations.

In order to consolidate their insights within an ontologically consistent framework free of misleading metaphors and category mistakes, we approached the problem from the radically different perspective of understanding *information preservation challenges* as *communication challenges*, appealing to established theories and paradigms for theorizing communication and representation, and building on previous research in information representation we conducted as Data Conservancy Data Concepts (DCDC) research group at the University of Illinois.

Starting from the Basic Representation Model (BRM) [Wickett et al., 2012], the notions of

situation and *situation type* from Situation Theory [Devlin, 1995], along with an account of *informing* derived from the Gricean theory of meaning, have been adopted to develop a conceptual framework of information communication useful to analyze information preservation through a communication lens.

This conceptual framework is then applied to analyze information preservation scenarios where artefacts are intended to carry information and to support its access over time.

While a *concrete representation* is always required in order for an agent to be presented with certain information —and therefore *material preservation* applies to a certain extent—, *what* information the agent is presented with is a function of the *interpretive frame* [Dubin et al., 2011] the agent brings into bear in a contextual cognitive interpretation process.

We distinguished between *persistent* material artefacts functioning directly as *concrete representations* of information for human access, and information intended to be carried by *ephemeral*, yet concrete, performances of digital materials.

We emphasized that a different level of communication needs to happen at the machine level to produce such performances. This communication involves interpreting bit–level representations according to appropriate *machine–level interpretive frames* and needs to be:

1. reliable —i.e. it needs to happen in virtue of established procedures appropriately executed;

2. sustained — i.e. it can be repeated over time and across changes in the underlying technology.

When communication is reliable and sustained, performances of a particular intended *type*—i.e. performances that exemplify the essential characteristics encoded by an intended digital material— can be produced over time and across changes in the underlying technology.

When a performance of a particular intended type is produced, a user attuned to an appropriate *user–level interpretive frame* is in the position to be presented with the information

that performances of that type are intended to carry.

From a digital stewardship perspective, 'preserving information' is therefore understood as

> *successfully sustained reliable communication through time and across changes in the mediating digital technology*

where both levels of communication —the machine and the human one— need to be sustained by means of capturing and documenting the agents' requirements, in terms of relevant levels of representations and appropriate interpretive frames, to successfully resolve an intended communication act. The *appropriateness* here is a function of decisions made at the time of 'creation' of a particular digital material according to a communication intent and possible transformative preservation actions applied according to particular preservation intents.

8.1 Future work

The research presented here is intended to move forward the agenda of defining sound conceptual foundations for digital stewardship by tackling the problem of *preserving information*.

Other aspects related to the informative function of, and experiential engagement with, digital materials can benefit from a similar attention and analytical treatment.

As we noted above, the communication of information is a highly interactive process, where many contextual factors affect *what* information emerges. For example, when scientific data are reused in contexts other than the one where the data were generated, other information might emerge in virtue of potentially different interpretive frames applied in the process and different contextual information. The problem of capturing contextual information necessary for the application of an appropriate interpretive frame was extensively discussed within the digital stewardship community [CCSDS, 2002, Lee, 2011]. Perspectives from Situation

Semantics can be applied to help understand and model how different contextual information can inflect what digital materials carry for users of different communities.

The conceptual framework we developed here already addressed certain experiential aspects of the user engagement with digital materials. However, certain digital materials exhibit *dynamic and interactive components*, where the users' input actively affects the *experiential engagement* with those materials. Example in this sense are complex digital objects that are database–driven —where query from users affects the characteristics of performances produced over time [Duranti and Thibodeau, 2006, Duranti et al., 2008]— or interactive media like video games [McDonough, 2011, McDonough, 2012, Sacchi and McDonough, 2012]—where each playing session can lead to very different performances as well. The question of what it really means to preserve certain interactive characteristics, what such a preservation entails, and what is their relations with those characteristics intended to carry information, certainly require extensions to the theory presented here.

While we recognise that this research is constrained to the problem of preserving information and does not cover the entire spectrum of characteristics that might be required for digital stewardship to be considered successful, the work presented in this dissertation sets the stage to continue research in these areas by showing how the analysis of fundamental concepts in digital stewardship can move forward the conceptual foundations of digital stewardship agenda and inform future practice.

REFERENCES

[Barateiro et al., 2010] Barateiro, J., Antunes, G., Freitas, F., and Borbinha, J. (2010). Designing digital preservation solutions: A risk management-based approach. *International Journal of Digital Curation*, 5(1):4–17.

[Barwise, 1986] Barwise, J. (1986). Information and circumstance. *Notre Dame Journal of Formal Logic*, 27(3):324–338.

[Barwise and Perry, 1980] Barwise, J. and Perry, J. (1980). *The situation underground*. Stanford University Press.

[Becker et al., 2009] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., and Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, 10(4):133–157.

[Boston, 1998] Boston, G. (1998). *Safeguarding the documentary heritage: a guide to standards, recommended practices and reference literature related to the preservation of documents of all kinds*. United Nations Educational, Scientific and Cultural Organization.

[Buckland, 1997] Buckland, M. (1997). What is a "document"? *JASIS*, 48(9):804–809.

[Capurro and Hjørland, 2005] Capurro, R. and Hjørland, B. (2005). The concept of information. *Annual review of information science and technology*, 37(1):343–411.

[Casati and Varzi, 2010] Casati, R. and Varzi, A. (2010). Events. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2010 edition.

[CCSDS, 2002] CCSDS, J. (2002). Reference model for an open archival information system (OAIS). Technical report, CCSDS 650.0-B-1, Blue Book.

[Chen, 2001] Chen, S. S. (2001). The paradox of digital preservation. *Computer*, 34(3):2428.

[Chowdhury, 2010] Chowdhury, G. (2010). From digital libraries to digital preservation research: the importance of users and context. *Journal of documentation*, 66(2):207–223.

[Compton et al., 2012] Compton, M., Barnaghi, P., Bermudez, L., Garcia-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al. (2012). The ssn ontology of the w3c semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*.

[Dappert and Farquhar, 2009] Dappert, A. and Farquhar, A. (2009). Significance is in the eye of the stakeholder. In *Research and Advanced Technology for Digital Libraries*, pages 297–308. Springer.

[Devlin, 2006] Devlin, K. (2006). Situation theory and situation semantics. *Handbook of the History of Logic*, 7:601664.

[Devlin, 1995] Devlin, K. J. (1995). *Logic and information*. Cambridge University Press.

[Dretske, 1983] Dretske, F. (1983). *Knowledge and the Flow of Information*.

[Dubin et al., 2009] Dubin, D., Futrelle, J., Plutchak, J., and Eke, J. (2009). Preserving meaning, not just objects: semantics and digital preservation. *Library Trends*, 57(3):595–610.

[Dubin et al., 2011] Dubin, D., Wickett, K. M., and Sacchi, S. (2011). Content, format, and interpretation. In Usdin, B. T., editor, *Proceedings of Balisage: the Markup Conference 2011*, volume 7 of *Balisage Series on Markup Technologies*, Montreal, Canada.

[Duerr et al., 2004] Duerr, R., Parsons, M., Marquis, M., Dichtl, R., and Mullins, T. (2004). Challenges in long-term data stewardship. In *21st IEEE Conference on Mass Storage Systems and Technologies, NASA/CP-2004*, volume 212750, pages 47–67. Citeseer.

[Duranti, 2005] Duranti, L. (2005). The interpares project: the long-term preservation of authentic electronic records: the findings of the interpares project.

[Duranti et al., 2008] Duranti, L., Preston, R., and Italiana, A. N. A. (2008). *International research on permanent authentic records in electronic systems (InterPARES) 2: Experiential, interactive and dynamic records*. CLEUP.

[Duranti and Thibodeau, 2006] Duranti, L. and Thibodeau, K. (2006). The concept of record in interactive, experiential and dynamic environments: the view of interpares*. *Archival Science*, 6(1):13–68.

[Furner, 2004a] Furner, J. (2004a). Conceptual analysis: A method for understanding information as evidence, and evidence as information. *Archival science*, 4(3/4):233–265.

[Furner, 2004b] Furner, J. (2004b). Information studies without information. *Library Trends*, 52(3):427446.

[Gangemi et al., 2002] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with DOLCE. *Knowledge engineering and knowledge management: Ontologies and the semantic Web*, pages 223–233.

[Giaretta, 2008] Giaretta, D. (2008). The CASPAR approach to digital preservation. *International Journal of Digital Curation*, 2(1):112–121.

[Giaretta et al., 2009] Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S., Guercio, M., Michetti, G., and Sawyer, D. (2009). Significant properties, authenticity, provenance, representation information and OAIS. *Proceedings of iPRES*, pages 67–73.

[Grice, 1957] Grice, H. (1957). Meaning. *The philosophical review*, pages 377–388.

[Grice, 1968] Grice, H. (1968). Utterer's meaning, sentence-meaning, and word-meaning. *Foundations of Language*, pages 225–242.

[Grice, 1969] Grice, H. P. (1969). Utterer's meaning and intention. *The philosophical review*, pages 147–177.

[Gruber et al., 1993] Gruber, T. et al. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

[Guarino, 1995] Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human Computer Studies*, 43(5):625–640.

[Guarino, 2002] Guarino, N. (2002). Ontology-driven conceptual modelling. In *Proc. of the 21st International Conference on Conceptual Modeling, LNCS*, volume 2503.

[Guarino and Welty, 2000] Guarino, N. and Welty, C. (2000). A formal ontology of properties. *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, page 191230.

[Guarino and Welty, 2002] Guarino, N. and Welty, C. (2002). Evaluating ontological decisions with ontoclean. *Communications of the ACM*, 45(2):61–65.

[Guizzardi et al., 2003] Guizzardi, G., Herre, H., and Wagner, G. (2003). On the general ontological foundations of conceptual modeling. *Conceptual ModelingER 2002*, pages 65–78.

[Haslanger and Kurtz, 2006] Haslanger, S. and Kurtz, R. (2006). *Persistence: contemporary readings*. The MIT Press.

[Hedstrom, 1997] Hedstrom, M. (1997). Digital preservation: a time bomb for digital libraries. *Computers and the Humanities*, 31(3):189202.

[Hedstrom and Lee, 2002] Hedstrom, M. and Lee, C. A. (2002). Significant properties of digital objects: definitions, applications, implications. In *Proceedings of the DLM-Forum*, page 21827.

[Heslop et al., 2002] Heslop, H., Davis, S., Wilson, A., and Australia, N. A. o. (2002). An approach to the preservation of digital records. Technical report, National Archives of Australia Canberra.

[Hofweber, 2013] Hofweber, T. (2013). Logic and ontology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2013 edition.

[IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998] IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional requirements for bibliographic records: final report*, volume 19. KG Saur Verlag Gmbh & Co.

[InterPARES, 2001] InterPARES (2001). *The Long-term Preservation of Authentic Electronic Records*, chapter Preservation Task Force Report. University of British Columbia.

[InterPARES, 2013] InterPARES (2013). Preservation.

[Knight, 2009] Knight, G. (2009). Inspect framework report.

[Kuny, 1998] Kuny, T. (1998). The digital dark ages? challenges in the preservation of electronic information. *International Preservation News*, page 813.

[Lee, 2011] Lee, C. A. (2011). A framework for contextual information in digital collections. *Journal of Documentation*, 67(1):95–143.

[Lee, 2012] Lee, C. A. (2012). *Handbook of Technical Communication*, chapter Digital Curation as Communication Mediation. De Gruyter Mouton.

[Lee et al., 2002] Lee, K. H., Slattery, O., Lu, R., Tang, X., and McCrary, V. (2002). The state of the art and practice in digital preservation. *Journal of Research-National Institute of Standards and Technology*, 107(1):93106.

[Lynch, 2013] Lynch, C. (2013). Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust. *Museums in a Digital Age*, page 314.

[Madin et al., 2007] Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279296.

[McDonough, 2011] McDonough, J. (2011). Packaging videogames for long-term preservation: Integrating FRBR and the OAIS reference model. *Journal of the American Society for Information Science and Technology*, 62(1):171–184.

[McDonough, 2012] McDonough, J. (2012). 'Knee-Deep in the Data': Practical Problems in Applying the OAIS Reference Model to the Preservation of Computer Games. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 1625–1634. IEEE.

[McGrath, 2012] McGrath, M. (2012). Propositions. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2012 edition.

[Mois et al., 2009] Mois, M., Klas, C.-P., and Hemmje, M. L. (2009). Digital preservation as communication with the future. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–8. IEEE.

[Moore, 2008] Moore, R. (2008). Towards a theory of digital preservation. *International Journal of Digital Curation*, 3(1).

[Pearce-Moses, 2012] Pearce-Moses, R. (2012). Preservation.

[Phelan, 2012] Phelan, P. (2012). *Unmarked: The politics of performance*. Routledge.

[Pike, 1954] Pike, K. L. (1954). Language in relation to a unified theory of the structure of human behavior.

[Reitz, 2004a] Reitz, J. M. (2004a). Digital preservation.

[Reitz, 2004b] Reitz, J. M. (2004b). Preservation.

[Renear and Dubin, 2007] Renear, A. and Dubin, D. (2007). Three of the four FRBR Group 1 entity types are roles, not types. *Proceedings of the American Society for Information Science and Technology*, 44(1):1–19.

[Renear and Choi, 2006] Renear, A. H. and Choi, Y. (2006). Modeling Our Understanding, Understanding Our Models–The Case of Inheritance in FRBR. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–16.

[Ross, 2012] Ross, S. (2012). Digital preservation, archival science and methodological foundations for digital libraries. *New Review of Information Networking*, 17(1):43–68.

[Sacchi and McDonough, 2012] Sacchi, S. and McDonough, J. (2012). Significant properties of complex digital artifacts: open issues from a video game case study. In *Proceedings of the 2012 iConference*, pages 572–573. ACM.

[Sacchi and Wickett, 2012] Sacchi, S. and Wickett, K. M. (2012). Taking modeling seriously [in digital curation]. In *Research Challenges in Digital Preservation - iPRES 2012*, Toronto, ON, Canada.

[Sacchi et al., 2011a] Sacchi, S., Wickett, K. M., Renear, A. H., and Dubin, D. (2011a). One thing is missing or two things are confused: An analysis of the OAIS Representation Information. In *7th Internation Digital Curation Conference (IDCC)*, Bristol, UK.

[Sacchi et al., 2011b] Sacchi, S., Wickett, K. M., Renear, A. H., and Dubin, D. S. (2011b). A framework for applying the concept of significant properties to datasets. In *Proceedings of ASIS&T 2011: the 74th Annual Meeting of the American Society for Information Science and Technology*, volume 48, New Orleans, LA.

[Sandore and Unsworth, 2010] Sandore, B. and Unsworth, J. (2010). *ECHO DEPository—Phase 2: 2008–2010 Final Report of Project Activities*, section 4.2.3, pages 33–37. University of Illinois, Champaign, IL.

[Strawson, 1963] Strawson, P. (1963). *Individuals: An essay in descriptive metaphysics*, volume 81. Taylor & Francis.

[Strodl et al., 2007] Strodl, S., Becker, C., Neumayer, R., and Rauber, A. (2007). How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 29–38. ACM.

[Svenonius, 2000] Svenonius, E. (2000). *The intellectual foundation of information organization*. MIT press.

[Task Force on Archiving of Digital Information et al., 1996] Task Force on Archiving of Digital Information, Commission on Preservation and Access, and Research Libraries Group (1996). *Preserving digital information : report of the Task Force on Archiving of Digital Information*. Commission on Preservation and Access, Washington, D.C.

[Thibodeau, 2002] Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. *The state of digital preservation: an international perspective*, page 431.

[Unsworth and Sandore, 2010] Unsworth, J. and Sandore, B. (2010). Echo depository-phase 2: 2008-2010 final report of project activities.

[Waugh et al., 2000] Waugh, A., Wilkinson, R., Hills, B., and Dell'Oro, J. (2000). Preserving digital information forever. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 175–184. ACM.

[Webb, 2003] Webb, C. (2003). *Guidelines for the preservation of digital heritage*. Information Society Division, United Nations Educational, Scientific and Cultural Organization.

[Wickett et al., 2012] Wickett, K. M., Sacchi, S., Dubin, D. S., and Renear, A. H. (2012). Identifying content and levels of representation in scientific data. In *Proceedings of ASIS&T 2012: the 75th Annual Meeting of the American Society for Information Science and Technology*, volume 49, Baltimore, MD.

[Yakel, 2007] Yakel, E. (2007). Digital curation. *OCLC Systems & Services*, 23(4):335–340.

[Zalta, 1983] Zalta, E. (1983). *Abstract objects: An introduction to axiomatic metaphysics*. Number 160. Springer Science & Business Media.