

Chapter 3

Introduction to OAIS Concepts and Terminology

If language is not correct, then what is said is not what is meant; if what is said is not what is meant, then what must be done remains undone; if this remains undone, morals and art will deteriorate; if justice goes astray, the people will stand about in helpless confusion. Hence there must be no arbitrariness in what is said. This matters above everything.

(Confucius)

This chapter aims to provide the basic ideas and concepts needed to build the rest of this book on. We do this by jumping in feet first, based on the terminology from the OAIS Reference Model. We need to do this in order to be able to talk clearly about digital preservation, because we want to say what we mean.

Another way of looking at this is to realise that different people have slightly different definitions in mind, depending upon their backgrounds, for many common terms. If we are not careful we will talk at cross-purposes because of these differences. In order to avoid this we need clear definitions.

The next few sections discuss some of the basic OAIS definitions and concepts.

3.1 Preserve What, for How Long and for Whom?



The “O” in OAIS stands for “Open” but refers to the open way the standard was developed rather than anything to do with open-access. Indeed the OAIS Reference Model can apply to any type of archive whether open access, closed, restricted, “dark” or proprietary.

OAIS takes a very general definition of its prime concern which, as the “T” in OAIS suggests, is information:

Information: *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.

Note that Knowledge is not defined in OAIS.

The accompanying definition of data is equally broad:

Data: *A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.* Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.

And in the case of things digital:

Digital Object: *An object composed of a set of bit sequences.*



Note that this does not mean we are restricted to a single file. The definition includes multiple, perhaps distributed, files, or indeed a set of network messages.

The restriction to “bits” i.e. consisting of “1” and “0”, means that if we move to trinary (i.e. “0”, “1” and “2”) instead of binary then we would have to change this definition, but it would not affect the concept – however it would change the tools we could use.

One might wonder why data includes physical objects such as a “moon rock specimen”. The answer should become clear later but in essence the answer is that to provide a logically complete solution to digital preservation one needs, eventually, to jump outside the digital, if only, for example, to read the label on the disk.

As to the question of length of time we need to be concerned about, OAIS provides the following pair of definitions (the text in bold italics below is taken from OAIS):

Long Term: *A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future.*

Long Term Preservation: *The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.*

In other words we are not only talking about decades into the future but, as is a common experience, we need to be concerned with the rapid change of hardware and software, the cycle time of which may be just a few years. Of course even if an archive is not itself looking after the digital objects over the long term, even by that definition, the intention may be for another archive to take over later. In this case the first archive needs to capture all the “metadata” needed so that it can hand these on also.

Three key concepts are embedded in the above definition namely:

Authenticity: *The degree to which a person (or system) may regard an object as what it is purported to be. The degree of Authenticity is judged on the basis of evidence.*

There will be much more to say about authenticity in Chap. 13, where the whole chapter is devoted to it.

Independently Understandable: *A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.*



By being able to “understand” a piece of information is meant that one can do something useful with it; it is not intended to mean that one understands all of its ramifications.

For example in a criminal investigation of a murder one may have a database with digitally encoded times of telephone calls; here we would be satisfied if we could say “the telephone call was made at 12:05 pm on 1st January 2009, UK time”, but to then understand that this implied that the person who made the call was the murderer is beyond what OAIS means by being able to “understand” the data.

Now we approach one element of what that the “preservation” part of “digital preservation” means. To require that things are able to be “interpreted, understood and used” is to make some very powerful demands. It not only includes playing a digital recording so it can be heard, or rendering an image or a document so that it can be seen; it also includes being able to understand what the columns in the spreadsheet we mention earlier means, or what the numbers in a piece of scientific data mean; this is needed in order to actually understand and in particular **use** the data, for example using it in some analysis programme, combining it with other data in order to derive new scientific insights. The “Independently” part is to exclude the easy but unreliable option of being able to simply ask the person who created the digital object; unreliable not because the creator may be a liar but rather because the creator may be, and in the very long term certainly will be, deceased!

Finally, we have the other key concept of “Designated Community”.

Designated Community: *An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the archive and this definition may change over time.*

Why is this a key concept? To answer that question we need to ask another fundamental question, namely “How can we tell whether a digital object has been successfully preserved?” – a question which can be asked repeatedly as time passes.

Clearly we can do the simple things like checking whether the bit sequences are unchanged over time, using one or more standard techniques such as digital digests [15]. However just having the bits is not enough. The demand for the ability for the object to be “interpreted, understood and used” is broader than that – and of course it can be tested.

But surely there is another qualification, for is it sensible to demand that *anyone* can “interpret, understand and use” the digital object – say a 4 year old child?

Clearly we need to be more specific. But how can such a group be specified, and indeed who should choose? This seems a daunting task – who could possibly be in a position to do that?

The answer that OAIS provides is a subtle one. The people who can should be able to “interpret, understand and use” the digital object, and whom we can use to test the success or otherwise of the “preservation”, are defined by the people who are doing the preservation.



The advantage of this definition is that it leads to something that can be tested. So if an archive claims “we are preserving this digital object for astronomers” we can then call in an astronomer to test that claim.

The disadvantage is that the preserver could choose a definition which makes life easy for him/her – what is to stop that? The answer is that there is nothing to prevent that but who would rely on such an archive?

As long as the archive’s definition is made clear then the person depositing the digital objects can decide whether this is acceptable. The success or failure of the archive in terms of digital objects being deposited will be determined by the market. Thus in order to succeed the archive will have to define its Designated Community(ies) appropriately.



Different archives, holding the same digital object may define their Designated Communities as being different. This will have implications for the amount and type of “metadata” which is needed by each archive.

As we will discuss later on, we need to be able to be a more specific, and we will see, in Chap. 7, how this can be done.

3.2 What “Metadata”, How Much “Metadata”?

One fundamental question to ask is “What ‘metadata’ do we need?” The problem with “metadata” is that it is so broad that people tend to have their own limited view. OAIS provides a more detailed breakdown. The first three broad categories are to

do with (1) understandability, (2) origins, context and restrictions and (3) the way in which the data and “metadata” are grouped together.

The reason for this separation is that given some digitally encoded information one can reasonably ask whether it is usable, which is dealt with by (1). This is a separate question to the one about where this digital object came from, dealt with by (2). Since there are many ways of associating these things it seems reasonable to want to separate consider (3) separately.



It could be argued that to understand a piece of data one needs to know its context. However the discussion about “Independently Understandable” in the previous section points out that OAIS does not require understanding of all the ramifications so this separation of context from understandability is reasonable, although it does not mean that all context is excluded from understandability since a piece of “metadata” may have several roles.

The packaging is something which is separate from the content. The next few sub-sections briefly introduce these different categories; they will each be discussed in much greater detail in separate chapters.

3.2.1 Understandability (*Representation Information*)

One type of “metadata” we can immediately identify is that which we need to “interpret, understand and use” the digitally encoded information. OAIS defines this as:

Representation Information: *The information that maps a Data Object into more meaningful concepts.* An example of Representation Information for a bit sequence which makes up a FITS file might consist of the FITS standard which defines the format plus a dictionary which defines the meaning of keywords in the file which are not part of the standard.

Figure 3.1 indicates that the Representation Information is used to interpret the Data Object in order to produce the Information Object – something which one can then understand and use.

The OAIS definition of **Information Object** is: *A Data Object together with its Representation Information.* This is a very broad definition.

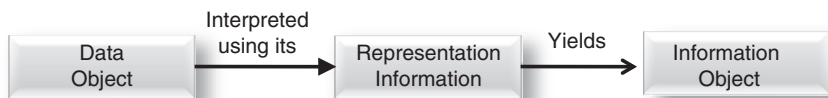


Fig. 3.1 Representation information



The definition of Information Object may seem a little circular. However its purpose is not to define something specific, for example in a computer programme. Instead it really only provides a simple term for something which we can apply to many different things in people's heads. The key idea is that it is something that allows us to talk about what knowledge is being exchanged.

When we are referring to something specifically targeted for preservation the term **Content Information** is used. This is *a set of information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information.*

In a little bit more detail, recognising that the Data Object could be either digital or physical, one can draw Fig. 3.2, which is a simple UML [257] diagram.

This diagram is a way of showing that

- an Information Object is made up of a Data Object and Representation Information
- a Data Object can be either a Physical Object or a Digital Object. An example of the former is a piece of paper or a rock sample.
- a Digital Object is made up of one or more Bits

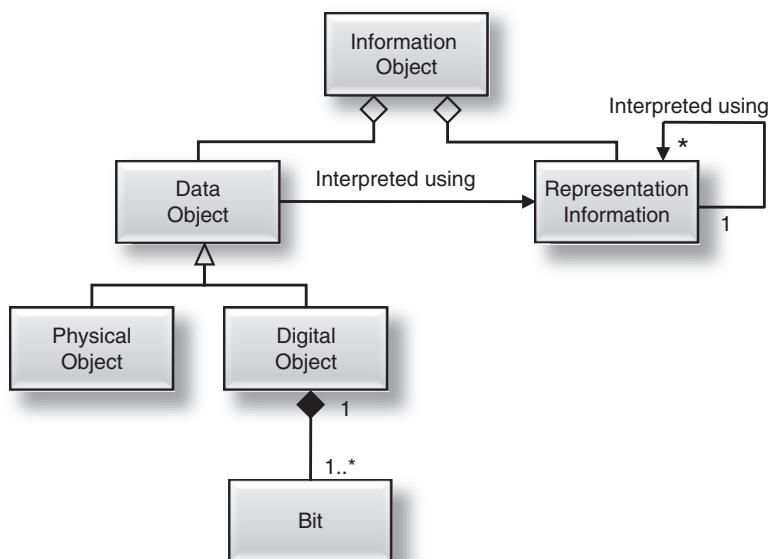


Fig. 3.2 OAIS information model



Note that this does not mean we are restricted to a single file. The definition includes multiple, perhaps distributed, files, or indeed a set of network messages.

- a Data Object is interpreted using Representation Information



It is important to realise that Representation Information can be anything from a scribbled handwritten note, needing a human to read it, to a complex machine readable formal description.

- Representation Information is itself interpreted using further Representation Information

Figure 3.3 denotes that Representation Information may usefully be sub-categorised into several different types, namely Structure, Semantic and (the imaginatively named) Other Representation Information. This breakdown is useful because Structure Representation Information is often referred to as “format”; Semantic Representation Information covers things such as ontologies and data dictionaries; Other Representation Information is a catch-all for anything and everything else.

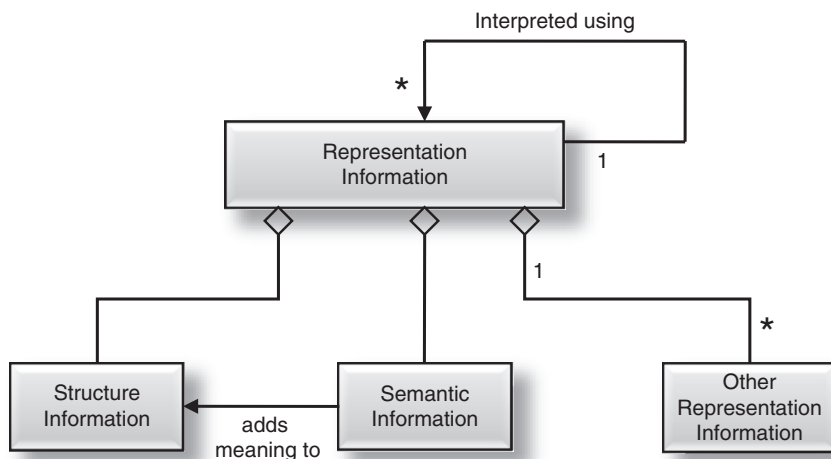


Fig. 3.3 Representation information object

One useful way to understand why this breakdown may be useful is to consider a number of different variations.



For example two copies of a simple message (i.e. a piece of information) may be contained in two text files (i.e. in the same format), but in one case the message is written in English and in the other case it is in French (needing different dictionaries).

Similarly one can have the English text both in a PDF and a Word file – two different formats but needing the same dictionary.

In general breaking things down into smaller pieces means that one is not forced to treat objects as a sticky mess. Instead one can deal with each (smaller) part separately and usually more easily.

When this is coupled with the fact that Representation Information is an Information Object that may have its own Data Object and other Representation Information associated with understanding that Data Object, as shown in a compact form by the *interpreted using* association, the resulting set of objects can be referred to as a Representation Network. Detailed examples will be provided in Part II.

In the extreme, the recursion of the Representation Information will ultimately stop at a physical object such as a printed document (ISO standard, informal standard, notes, publications etc). This allows us to make a connection to the non-digital world. However use of things like paper documentation would tend to prevent “automated use” and “interoperability”, and also complete resolution of the complete Representation Network, discussed further below, to this level would be an almost impossible task. Therefore we would prefer to stop earlier, and this will be discussed next.

As the final part of this rush through the OAIS concepts we turn to something a little different in order to answer the question “*How much ‘metadata’?*”

A piece of Representation Information is just another piece of Information – hence the name Representation Information rather than Representation Data. In order for there to be enough Representation Information it has to be understandable and usable by the Designated Community – in order to be used to understand the original data object. However what if this is not the case?

The Representation Information may be encoded as a physical object such as a paper document, or it may be a digital object. In the latter case we can simply provide Representation Information for that digital object. If the Designated Community still cannot understand and use the original data, we can repeat the process. Clearly this provides us with a way to answer the “How much” question: we provide a network of Representation Information until we have enough for the Designated Community to understand the Data Object. OAIS defines:

Representation Network: *The set of Representation Information that fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term.*

To complete the picture we can then see a way to define the Designated Community, namely we define them by what they know, by what OAIS terms their Knowledge Base:

Knowledge Base: *A set of information, incorporated by a person or system that allows that person or system to understand received information.*

All these terms will be discussed at much greater length in Chap. 6.

3.2.2 Origins, Context and Restrictions (Preservation Description Information)

OAIS defines a group of types of “metadata”, under the name of Preservation Description Information (PDI), which are to do, broadly, with knowing what and where the digital object came from.

The idea is that one needs to name a way to identify the digital object; to know how and by whom and why the digital object is what it is; to know the broader context within which it exists; to be sure that the digital object has not been changed and finally, to know what rights are attached to it (see Fig. 3.4).

The following sections provide the OAIS definitions with a little additional explanation; further details are provided in Chap. 10.

3.2.2.1 Reference Information

Reference Information: *The information that is used as an identifier for the Content Information. It also includes identifiers that allow outside systems to refer unambiguously to a particular Content Information.* An example of Reference Information is an ISBN. Clearly what are often called persistent identifiers, which we discuss further in Sect. 10.3.2, provide a form of Reference Information.

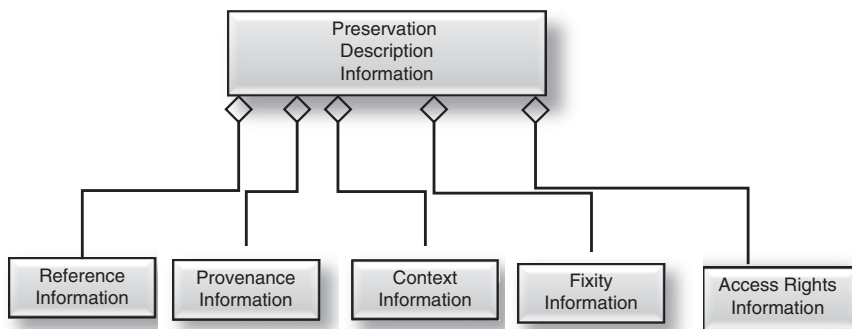


Fig. 3.4 Preservation description information

3.2.2.2 Provenance Information

Provenance Information: *The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. The archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity.*

Provenance may reasonably be divided into what we might term Technical Provenance – things that, for example, are recorded fairly automatically by software. This must be supplemented by Non-technical Provenance, by which we mean, for example, the information about the people who are in charge of the Content Information – the people who could perhaps fake the other PDI. In other words in order to judge whether we can trust the multitude of information that surrounds the Content Information, we must be able to judge whether we trust the people who were responsible for collecting it, and who may perhaps have been able to fake it. This will be discussed in more detail in Sect. 13.

3.2.2.3 Context Information

Content Information: *The information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects.*



It is worth noting here that many traditional archivists would say that “context” is all important and trumps all other considerations [16]. OAIS defines “context” in a rather more limited way, but on the other hands provides a greater level of granularity with which to work, although does point out that Provenance, for example, is a type of Context.

3.2.2.4 Fixity Information

Fixity Information: *The information which documents the mechanisms that ensure that the Content Information object has not been altered in an undocumented manner.* An example is a Cyclical Redundancy Check (CRC) code for a file.

Digests [15] are often used for this purpose, relying on the fact that a short bit sequence can be created, using one of several algorithms, from a larger binary object which it represents, essentially uniquely. By this we mean that it is, practically

speaking, impossible to design a different file with a matching digest. This means that if we can keep the (short) digest safely then we can use it to check whether a copy of a (perhaps very large) digital object is what we think it is. This can be done by recomputed the digest, using the same algorithm, using the digital object which we wish to check. If the digest matches the original one we carefully kept then we can be reasonably sure that the digital object does indeed have the same bit sequence as the original.

3.2.2.5 Access Rights Information

Access Rights Information: *The information that identifies the access restrictions pertaining to the Content Information, including the legal framework, licensing terms, and access control. It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer). It also includes the specifications for the application of rights enforcement measures.*

Access rights and digital rights are discussed further in Sects. 10.6 and 17.7.

Examples of PDI from different disciplines are given in Table 3.1.

3.2.3 Linking Data and “Metadata” (Packaging)

The idea behind packaging is that the one must somehow be able to bind the various digital objects together. Remember also that Content Information is the combination of Data Object plus Representation Information, and PDI has its various components. Fig. 3.5 shows the other conceptual components of a package.

The package does not need to be a single file – it is very important to understand this. It could be, but it does not have to be. The package is a logical construction, in other words one needs to be able to have something which leads one to the other pieces, by one means or another. About the package itself one needs to be able to identify it i.e. is it a file, a collection of files, a sequence of bytes on a tape? The information which provides this is the Packaging Information.

Packaging Information: *The information that is used to bind and identify the components of an Information Package.* For example, it may be the ISO 9660 volume and directory information used on a CD-ROM to provide the content of several files containing Content Information and Preservation Description Information. For a ZIP file it would be the information that the package is the file which probably has a name ending in “.zip”.

In addition the package contains something and the Package Description provides the description of what this is; it is something that can be used to search for this particular package.

Table 3.1 Examples of PDI

Content information type	Reference	Provenance	Context	Fixity	Access rights
Space science data	<ul style="list-style-type: none"> • Object identifier • Journal reference • Mission, instrument, title, attribute set 	<ul style="list-style-type: none"> • Instrument description • Principal investigator • Processing history • Storage and handling history • Sensor description • Instrument • Instrument mode • Deconvolution map • Software interface specification • Information property description 	<ul style="list-style-type: none"> • Calibration history • Related data sets • Mission • Funding history 	<ul style="list-style-type: none"> • CRC • Checksum • Reed-Solomon coding 	<ul style="list-style-type: none"> • Identification of the property authorized Designated community (access control) • Permission grants for preservation and for distribution • Pointers to fixity and provenance information (e.g., digital signatures, and rights holders)
Digital library collections	<ul style="list-style-type: none"> • Bibliographic description • Persistent identifier 	<ul style="list-style-type: none"> • For scanned collections: <ul style="list-style-type: none"> • “metadata” about the digitization process • pointer to master version • For born-digital publications: <ul style="list-style-type: none"> • pointer to the digital original • “Metadata” about the preservation process: <ul style="list-style-type: none"> • pointers to earlier versions of the collection item • change history • Information property description 	<ul style="list-style-type: none"> • Pointers to related documents in original environment at the time of publication 	<ul style="list-style-type: none"> • Digital signature • Checksum • Authenticity indicator 	<ul style="list-style-type: none"> • Legal framework(s) • Licensing offers • Specifications for rights enforcement measures applied at dissemination time • Permission grants for preservation and for distribution • Information about watermarking applied at submission and preservation time • Pointers to fixity and provenance information (e.g., digital signatures, and rights holders)

Table 3.1 (continued)

Content information type	Reference	Provenance	Context	Fixity	Access rights
Software package	<ul style="list-style-type: none">• Name• Author/originator• Version number• Serial number	<ul style="list-style-type: none">• Revision history• Registration• Copyright• Information property description	<ul style="list-style-type: none">• Help file• User guide• Related software• Language	<ul style="list-style-type: none">• Certificate• Checksum• Encryption• CRC	<ul style="list-style-type: none">• Designated community• Legal framework(s)• Licensing offers• Specifications for rights enforcement measures applied at dissemination time• Pointers to fixity and provenance information (e.g., digital signatures, and rights holders)

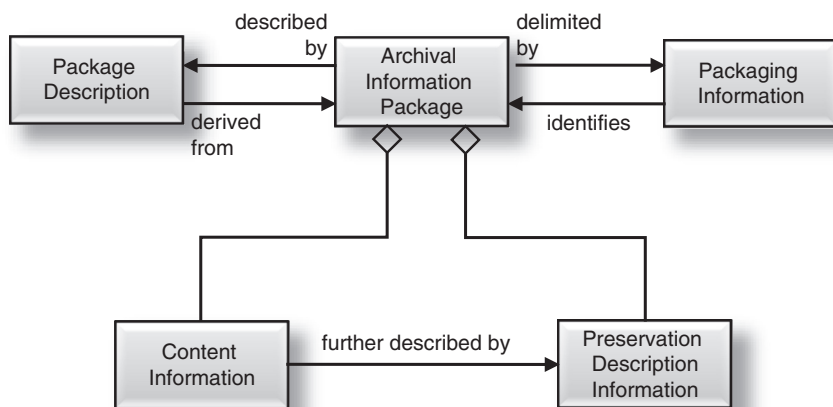


Fig. 3.5 Information package contents

It may perhaps have been noticed that the various additional concepts we have identified are called “Information”. In most cases these will be digitally encoded. This leads us to a fundamentally important point.

3.3 Recursion – A Pervasive Concept

Those with a mathematical background will recognise some of this as a type of recursion. It comes up time and again in preservation. By this we mean that ideas which appear at one level of granularity re-appear when we take a finer grained view, within the detailed breakdown of those or other ideas. As is well known in mathematics, it is important to understand where the recursion ends otherwise it becomes impossible to produce practical results. For example the factorial function is defined as $n! = n * (n-1)!$ i.e. $6! = 6 * (5!) = 6 * 5 * (4!) = \dots$ This stops when we get to $0!$ because we define $0!$ as equal to 1.



It is worth making some remarks about this concept here.

Representation Information (RepInfo for short) – remember it is Representation Information rather than Representation Data – is encoded as data (which could be called representation data but in fact OAIS does not use that terminology) which itself needs its own Representation Information. The recursion stops at the Knowledge Base of the Designated Community.

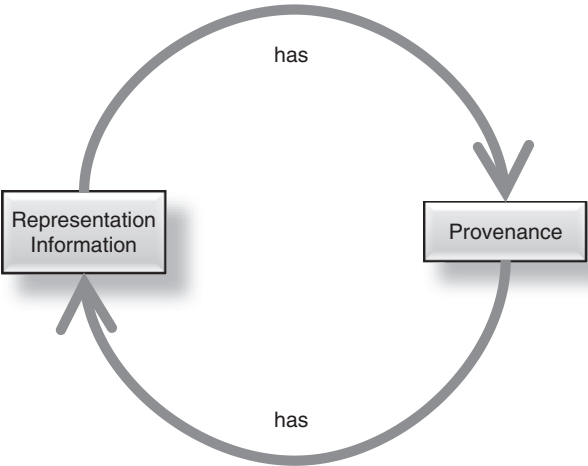


Fig. 3.6 Recursion – Representation information and provenance



Any piece of “metadata”, such as Provenance (to be discussed in detail later), will itself be encoded as a Data Object, which needs Representation Information. Representation Information as a digital object will also need its own Provenance, as illustrated in Fig. 3.6.

The recursion in this case might end with Provenance being a simple text file (or piece of paper) in plain English (assuming the Designated Community can read English) so the Representation Information is quite simple and hence the Representation Information Network terminates.

A formal way of showing this in OAIS is by showing that many of the concepts that are used are Information Objects as shown in Fig. 3.7.

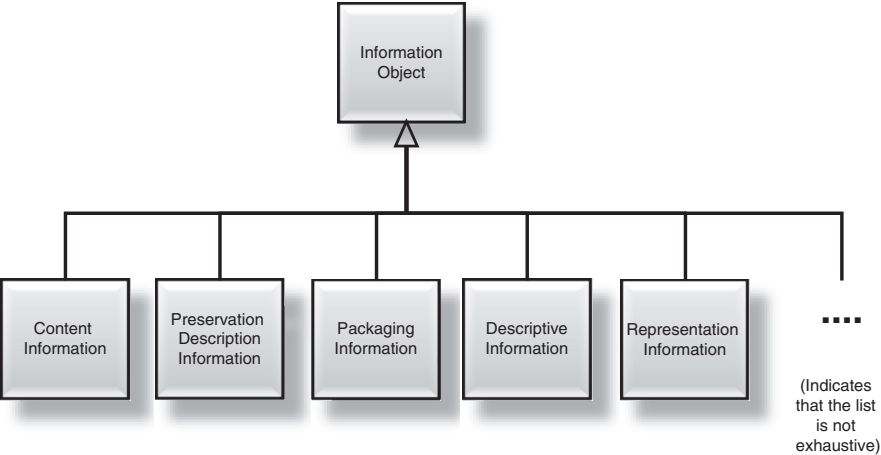


Fig. 3.7 Sub-types of information object

Components of a preservation infrastructure themselves need to be preservable – for example a Registry (see Sect. 16.2.1.1) which supports OAIS-type archives must itself be an OAIS-type archive in that it must be relied upon to preserve its (Representation) Information objects over the long term.

3.4 Disincentives Against Digital Preservation

It is important to realise that although many of those reading this book will regard preserving our digital heritage as self-evident, nevertheless this is not universal opinion.

As time passes more and more digitally encoded information is accumulated. It is therefore possible that the costs increase over time, yet experience tells us that the budget available for a preservation organisation usually does not. Figure 3.8 might therefore be projected to be the case.

If this is the projection then no responsible body would find it acceptable; a decision would have to be taken not to preserve everything – or perhaps not to preserve anything. The focus here is on how we could try to control the costs so that either the graph of preservation costs is level rather than increasing, or is increasing only slowly so that the crossing-point is acceptably far into the future.

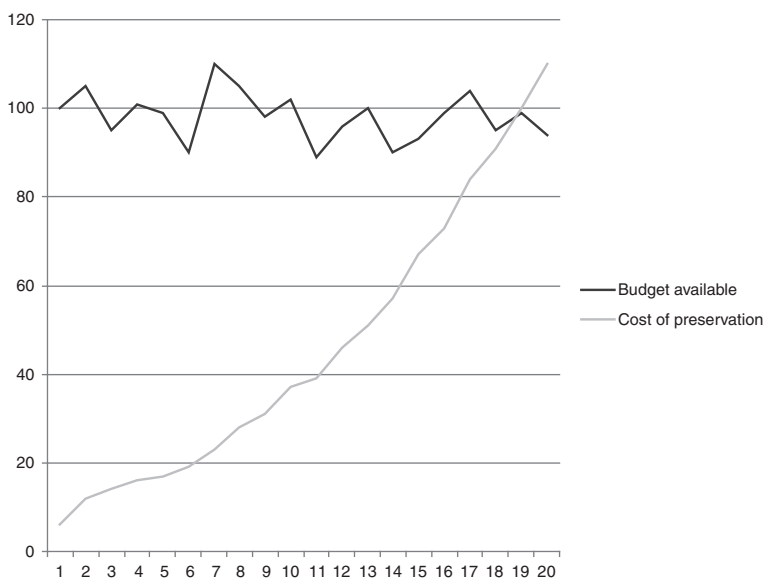


Fig. 3.8 Money disincentives – if the annual cost of preservation of the accumulated data increases over time

3.4.1 Cost/Benefit Modelling

It is very hard to model the costs of digital preservation [17], and even more difficult to evaluate possible benefits. However it is worth discussing at least some of the costs at this point to illustrate the point.

One of the simplest costs which one may try to estimate is that of storage. The argument sometimes used is that the cost of a unit of storage reduces by 50% each technology cycle (say 3 years). Suppose that the initial cost is £X. If at each cycle one buys new hardware then one spends £X/2 in 3 year's time, £X/4 after a further 3 years, and so on. Therefore one would spend

$$£X + £X/2 + £X/4 + £X/8 + \dots = £2X \text{ in total}$$

Thus one can argue that the hardware cost is at least controlled.

However each 3 years the amount of data may easily have increased by, say, a factor of 8, thus the cost keeping all the data would be:

$$\begin{array}{ccccccc} X & + & X/2 & + & X/4 & + \dots & = & 2*X \\ & & 8*X/2 & + & 8*X/4 & + \dots & & 8*X \\ & & & & 64*X/4 & + \dots & & 32*X \end{array}$$

Thus one can see that there is a real danger that the growth of data volumes may easily swamp the cost savings introduced by new technologies. Moreover the cost of personnel, and, more importantly, the cost of preserving the information rather than simply keeping the bits, has been left out of the calculations.

More complex modelling is available based on cost data from real, anonymised, archives [18]. However the cost models which are available omit the cost of maintaining understandability, which could be labour intensive.

The Blue Ribbon Task Force on Sustainable Preservation and Access [19] has looked at the broader view and identifies the fact that one can effectively purchase “future options” without making an indefinite commitment.

3.4.2 Future Generations

Although preserving our digital heritage for future generations is a laudable ambition, it must be admitted that those future generations have two great weaknesses (1) they do not (yet) have a vote and (2) they do not (yet) pay taxes! As a result other priorities can overwhelm our ambitions, no matter how laudable.

Clearly one cannot preserve everything and there are always more or less formal mechanisms to choose what to keep and what to leave to decay (or leave for someone else to preserve); it may be that the availability of funding determines what stays and what goes and, in the long term, if money runs out then the whole of the collection could die.

The ways of deciding what should be preserved is not part of this book in part because there are too many variations depending on particular circumstances. However there are a number of generally applicable techniques discussed in Chap. 14 concerning preservation objectives and building a business case for preserving digital assets.

3.5 Summary

In this chapter we have very briefly introduced the key concepts about digital preservation that will stand the reader in good stead throughout the rest of the book.



<http://www.springer.com/978-3-642-16808-6>

Advanced Digital Preservation

Giaretta, D.

2011, XXII, 510 p., Hardcover

ISBN: 978-3-642-16808-6