**OAIS Preservation Issues**

Donald Sawyer

15 May 2018

**1. Introduction**

The primary objective for the development of the OAIS Reference Model was to provide a reference framework of sufficient terms and concepts to facilitate communication among a wide variety of groups and individuals regarding the Long Term preservation of digital information. This has certainly been achieved and can only be considered extremely successful. However for the last year I’ve been concerned that there are aspects of the OAIS Reference Model that, from the beginning, have been too narrowly focused. I say this as one of the co-editors of the original version, and thus I feel a special responsibility for this situation. I have mentioned some of my concerns to some of the DAI participants who have a long history with OAIS. However until now I have not been able to fully define my concerns. This paper is my attempt to clearly state my concerns in a way that I hope will be understandable to the DAI participants. I realize that most of the formal comments on OAIS have already been resolved and therefore this input is late. However my concern is that a significant critique of OAIS may be published in the near future and I would like this to be avoided if possible. My hope is that whatever consensus is reached can be incorporated into the current update to OAIS.

The approach taken is to identify three topics for discussion and then provide conclusions based on the views put forth. The intent is to reach some consensus on the topics raised and not to propose specific updates to OAIS as the latter would be premature and likely counter productive.

The three topics for discussion are given as:

* The Definition of Information and the Importance of Information Presentation
* Information Modeling
* Role for Information Access and Use

In this document, terms capitalized correspond with their OAIS definitions. OAIS terms that are not capitalized mean these terms are used in ways that may, or may not, be at variance with current OAIS meaning.

**2. The Definition of Information and the Importance of Information Presentation**

**2.1 Discussion**

If a person wanders into a gallery and spots a black and white photograph the person has never seen before, and it makes a lasting impression, the person has new knowledge. It seems clear that the person’s base of knowledge has increased (image of picture, and the overall experience retained) by the presentation of this image to the person’s visual sense along with its subsequent retention. The experience has meant something to the person. Generalizing, it can be said that the presentation of signals perceptible to human senses (e.g. viewing the image) is the presentation of information that can result in an addition to the person’s knowledge base. This is important because, for the example above, what the person retains from the experience depends almost entirely on the person’s past experience and knowledge, and not on any specific meaning (there may have been none) that was intended to be conveyed by the image creator.

In other words, it would be problematic to say that the person now ‘understands’ the image where ‘understand’ has its common meaning and usual dictionary definition such as “to perceive the intended meaning”. It is meaningful to the person, but it cannot be said to be an ‘intended meaning from another person’. In this case the role of the knowledge base is not to understand (an intended meaning of) the new information, but to serve as the repository into which the new information (the experience) is to be integrated. This is an important distinction.

However, if the photo is accompanied by a description written in a natural language, it is expected that the person who views the description may be able to understand this information by making use of the person’s knowledge of the natural language. It follows that when a person perceives a signal, it is information that may be retained as an experience through ***addition*** to the person’s knowledge base, but that information may, or may not, also be intended to be ***understood*** by the person’s knowledge base.

In the example above, the information received by viewing the image caused the person’s knowledge base to be incremented with this experience, and therefor it would be logical to conclude that the information was some type of knowledge. Surely a person’s knowledge base includes retention of past experiences. However given our OAIS Information definition as “any type of knowledge that can be exchanged”, I was not thinking about information that was simply to result in an experience. In fact this type of information could be ruled out given the longstanding OAIS definition of Knowledge Base as “A set of Information, incorporated by a person, that allows that person to understand the received Information”. For example, given that it is perfectly reasonable that an Archive should be able to preserve a digital recording of a musical performance, and that the objective of this preservation is most likely to maintain the ability to recreate the presentation to provide an experience of the performance, it has to be concluded that our OAIS definitions of Information and Knowledge Base, taken together, are inadequate. It may be argued that one could try to adopt, as a minimum, a unique and very broad meaning for ‘understand’. Even if this could be done, it would contravene everyone’s common understanding of the term and the fundamental purpose of the OAIS Reference Model to promote clear communication about digital preservation.

Our definition of Information was taken from an ISO standard that also discussed ‘Domain of Knowledge’, and if memory serves, ‘Knowledge Base’, and it lead me to think that the Information so identified had come from someone who understands something and that the intent in the exchange is to convey that understanding (meaning) to the recipient person. (I’m ignoring computer proxy systems sending and/or receiving information for now.) This is certainly how I thought about it at the time we developed OAIS, and I believe most others did as well. The intent to preserve meaning encoded in digital information is certainly critical and OAIS has rightly made this concern foremost. This has to be a major reason for the success of OAIS to this point. After all, one cannot get much meaning from a sequence of bits. But as has been shown, this focus only on understanding is too narrow. Our senses are receiving and experiencing new or redundant Information all the time. The signals that our senses are receiving qualify as OAIS Data, but they are also information. (For a perspective on the state of information theory, and some open issues, see http://philsci-archive.pitt.edu/10911/1/What\_is\_Shannon\_Information.pdf. Some view the Universe as simply information, or just mathematics. See “Our Mathematical Universe: My Quest for the Ultimate Nature of Reality” by Max Tegmark.) Note that the signals (Information) mentioned above are not OAIS Information Objects because they are not accompanied by Representation Information. However one could think of the Representation Information as being incorporated in the physical (sensory) makeup, including knowledge base, of each individual. Of course each individual’s experience resulting from exposure to the same information (signals) will be somewhat unique.

My conclusion from the above, as it applies to the current OAIS Reference Model and draft update to the OAIS Reference Model, is that our current definitions of Information and Knowledge Base, given as:

* Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by data.; and
* Knowledge Base: A set of Information, incorporated by a person or their proxy system that allows that person or their proxy system to understand the received information (highlighted text in draft version);

taken together are too narrow and thus too limiting. One approach to addressing this issue would be to keep the Information definition, but update the Knowledge Base definition to make clear that ‘experience’ as well as ‘understand’ is included. However this is not a proposal and is made to help clarify the concern.

These definitions, which I had a major hand in promoting, have also led me to significantly underestimate the importance of the presentation of information to human senses in the overall preservation process. Not only is presentation critical for Information that is expected to only be experienced, it is also mandatory for all Information that is to be understood by a person. This follows because we perceive Information only when it is presented to one or more of the human senses, and it cannot be understood until it has been presented. I’ve come to believe that the ‘presentation of information’ should play a much more prominent role in the OAIS Reference Model.

In the current draft update to the OAIS Reference Model, there is mention of ‘proxy systems’ and also ‘client systems’. Using the term ‘proxy system’ as a stand-in for a person, one can ask to what extent the above discussion applies to such systems. A proxy system can be said to have some type of knowledge base as it responds to inputs and generates outputs. A proxy system that changes state in response to an input has clearly responded to that information. If it does not retain that information in some form, it would be logical to conclude that its knowledge base has not been updated and thus it has not retained the experience. Even so, it could still respond using its static knowledge base to transform the input to some corresponding output. However until the output is in a form presentable to human senses, it will not be available for human experience and possible understanding.

When information is created, or captured, and is intended to be shared, it is done with some purpose, or purposes, in the mind of the creator. From the perspective of preservation, an important dimension for the intended purpose can be to categorize it as: experience, understanding, or both experience and understanding. However when such information is presented for preservation, the Archive may assign a preservation objective that is at variance with, and often more limited than, the creator’s intended purpose. The Archive may adopt a preservation objective only to maintain the experience (e.g., historical document), or it may adopt a preservation objective to only maintain the understanding (meaning) (e.g. research report allowing very different presentation looks), or it may adopt a preservation objective to maintain both the experience and the understanding. This last case implies the need to associate other information that aids in maintaining understanding over the Long Term without altering the presentation of the original information.

It seems clear that concepts of Transformational Information Property and Designated Community are applicable regardless of whether the original intent for creating (or capturing) the information is to provide experience or understanding or both.

The perspectives presented in this section are focused on the role of Archives in preserving information and apply to all Archives. They are not addressed to issues surrounding the role of an Archive in supporting local requirements for ease of access and use by its Consumers. This is a separate topic and is addressed in section 4.

**2.2 Conclusions**

1. Current combined definitions of Information and Knowledge Base are too limiting as together they imply an OAIS does not include Information that is only for the purpose of eliciting an experience. Currently it must be for the purpose of understanding.
2. It should be recognized that information, needing to be preserved, has an intended purpose that arises from its creator (or capturer) and that can be categorized as: experience, understanding, or experience and understanding.
3. Information that is intended to be understood must also be presentable and therefore will, most likely, elicit an experience (updated knowledge) in a human, but not necessarily in a proxy system.
4. Information that is intended only to elicit an experience will mean the Archive will have a preservation objective of, at least, maintaining the ability to generate the signals (information) expected to elicit that experience.
5. Information that is intended to be both experienced and understood may have an Archive preservation objective more limited than the intended purpose (e.g., experience only).
6. Information that is preserved should be explicitly designated with one of the following preservation objectives: ‘experience’, ‘understanding’, or ‘experience and understanding’
7. Both information intended to elicit an experience and information intended to be understood must have explicit or implicit Transformational Information Properties that must be preserved in any Transformation.
8. Information, intended to be understood, may have two different Designated Community definitions, one for the experience (e.g., can’t be color blind, normal hearing) and one for the understanding (e.g., education level, language proficiency). They could be combined into one definition.
9. Because information can be neither experienced nor understood until it has been presented in a manner accessible to human senses, it should be incumbent on an Archive to maintain the appropriate presentation capability, or at least demonstrate how that can be achieved at any point in time. This provides a level of assurance that preservation is being maintained and it facilitates comparisons with separately maintained presentation examples (including required information properties) from the time of Ingest.

**3. Information Modeling**

**3.1 Discussion**

The OAIS Reference Model provides Information models in OAIS Reference Model section 2.2, which is the Information modeling that others are to use for OAIS conformance as addressed in OAIS Reference Model section 1.4. More detailed modeling is provided in OAIS Reference Model section 4.2. This modeling in not fully consistent with the discussion above in section 2. Figure 3-1 and 3-2 consist of several models and are provided to clarify these issues.

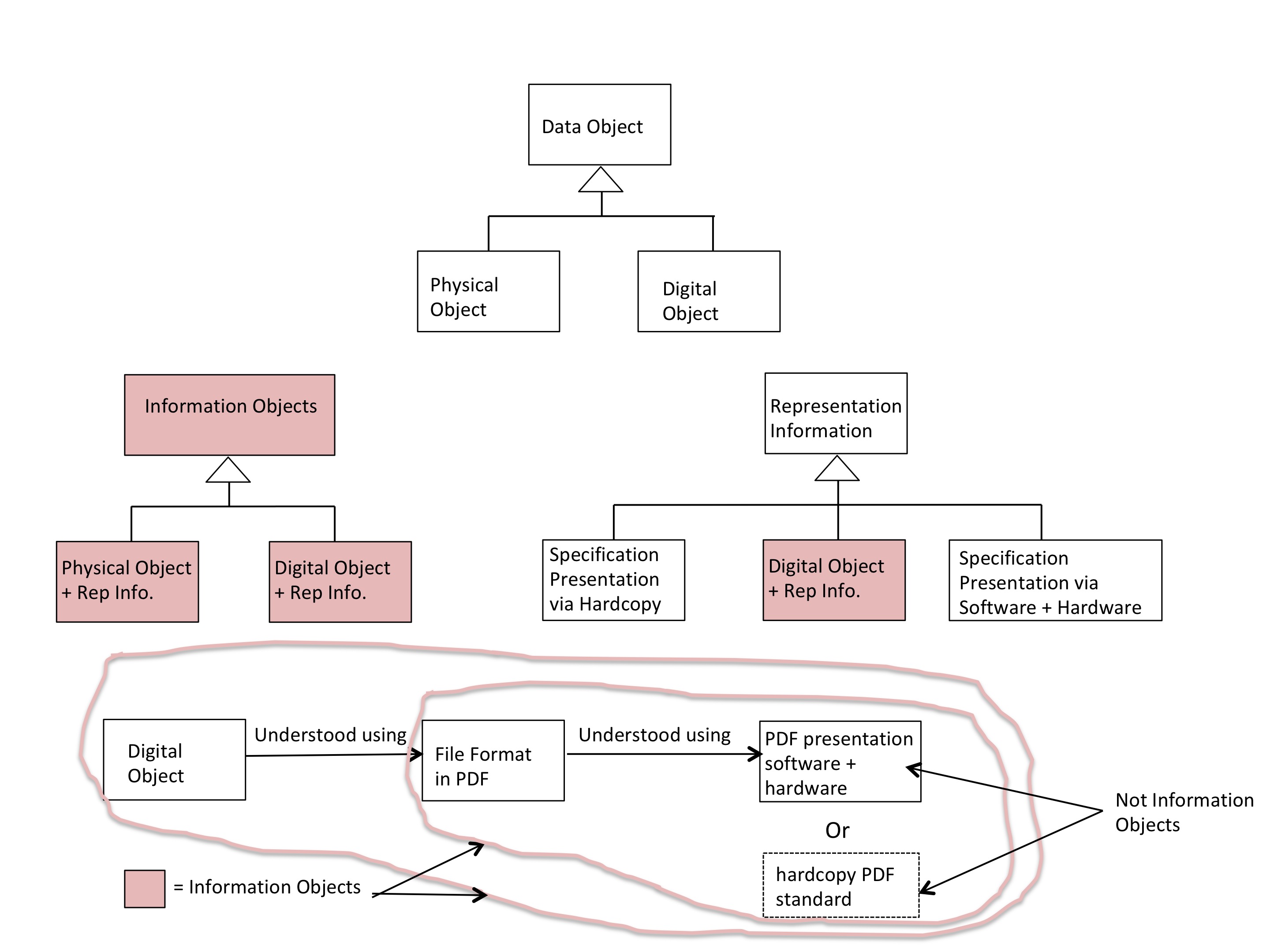


Figure 3-1: Some Information Models Consistent with Section 2

In Figure 3-1, the upper model shows a Data Object can be classified as either a Physical Object or a Digital Object as given in OAIS Terminology. The next model below and to the left shows an Information Object can be classified as either a Physical Object with Representation Information, or a Digital Object with Representation Information as also given in OAIS Terminology. To the right of this model is a model of Representation Information. Unlike in OAIS, Representation Information is not always an Information Object and when there is recursion forming a Representation Network, the leaf nodes must be presentations to human senses (e.g., via hardcopy or a software and hardware system) to stop the recursion. Note that software and hardware that is presenting a specification is not considered to be an Information Object any more than a hardcopy document presenting a specification is an Information Object. One could view the software as having the specification’s Representation Information incorporated into its knowledge base, just as one can view the human ability to make sense of sensory signals by virtue of Representation Information that is incorporated into the human sensory system and knowledge base, as discussed earlier. The model at the bottom of Figure 3-1 shows simple examples of an expansion of an Information Object where the Representation Network is terminated, as it must be, with a presentation of information via a physical device.

It has been argued that an Archive should be required to be able to demonstrate that it can present preserved information so that it is perceptible to human senses. Figure 3-2 shows examples of Information presentations that an Archive could use.

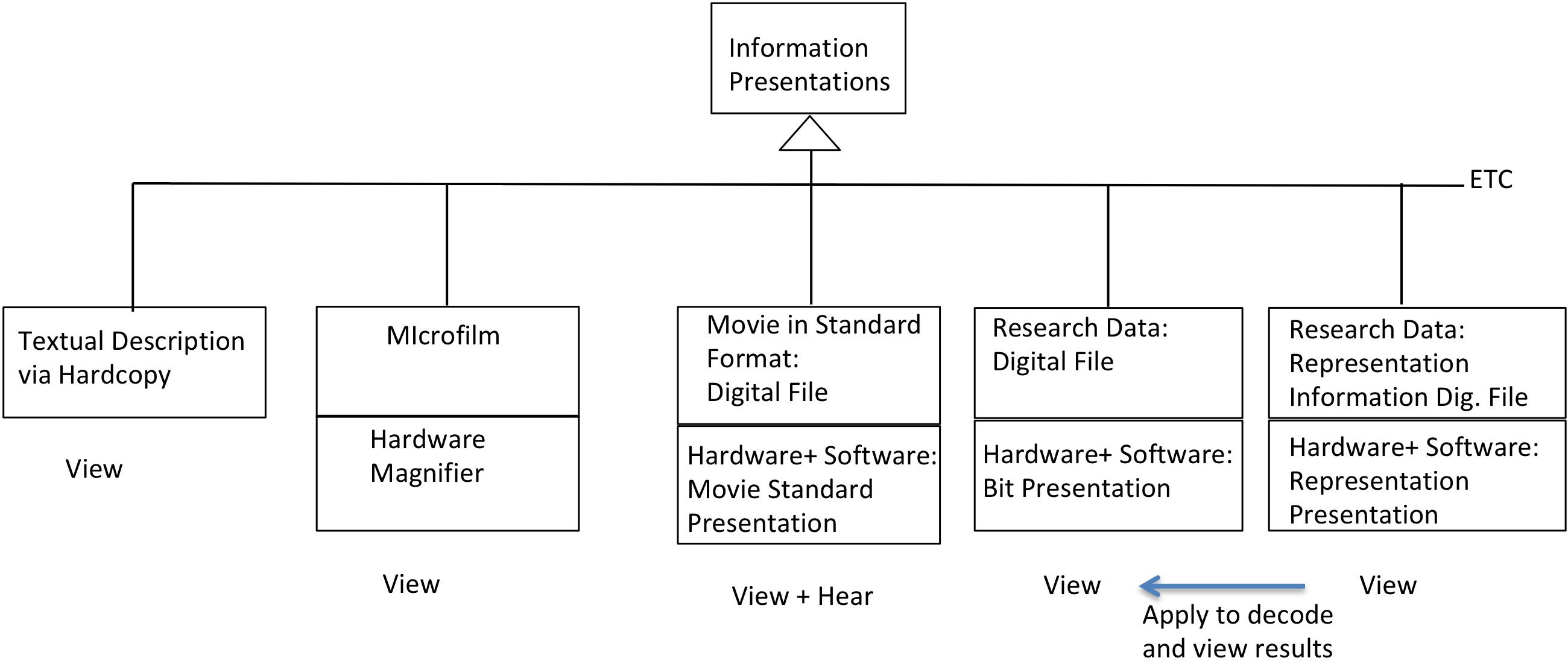


Figure 3-2: Some Information Presentation Categories

Starting from the left, it shows:

1. A hardcopy document written in a natural language is directly perceptible to the human visual sense and therefore does not need Representation Information.
2. The next example shows a sheet of microfilm that contains information that is directly perceptible to the human visual sense, however the human visual sense is not easily able to perceive the full information content without the use of a hardware magnification capability.
3. The third example shows a digital movie file encoded in a widely available standard movie format. A digital file it is not directly perceptible to human senses. Its Representation Information has been incorporated into software that, along with compatible hardware, is able to present the movie so that it is perceptible to the human senses of sight and hearing. It might be said that the Representation Information carried by the standard (or at least enough to show the movie in this system) has been incorporated into the knowledge base of the software. One cannot conclude that the software is a full replacement for the Representation Information as specified in the standard.
4. The fourth example shows a digital (encoded) file that contains research results. Software and hardware is used to present the bit information in a form that is perceptible to human vision, perhaps as 8-bit characters. However without the Representation Information, shown in the next example, the full information content is not perceptible.
5. The Representation Information for the file of example 4 is shown in the fifth example. The presentation of this information is carried out using software that understands the format of the digital Representation Information together with compatible hardware. The result is a visual presentation of the Representation Information. Also shown is the possibility of a person taking an understanding of the displayed Representation Information and manually applying it to the displayed bit information of example 4. This allows the person to decode the Research File and obtain full information. Alternatively, software could be written to replace the manual application and display the information efficiently.

Figure 3-3 expands on the decoding of digitally encoded information with examples showing three possibilities for the involvement of digital and physical Representation Information. In section A, a digital Representation component employing a non-widely used standard to decode the Digital Object is described by widely used digital standards and terminated in a hardcopy version of the UNICODE standard. Other options for the widely used standards include the use of a hardware and software combination to decode the digital object to which they are associated and to present the resulting Information.

In section B, the non-widely used standard is incorporated into software and hardware that is able to present the information encoded in the Digital Object. Maintaining preservation for this case requires maintaining this non-widely used software and hardware combination, or some facsimile of this system, or if practical, transforming the presented information of this system into a more maintainable form.

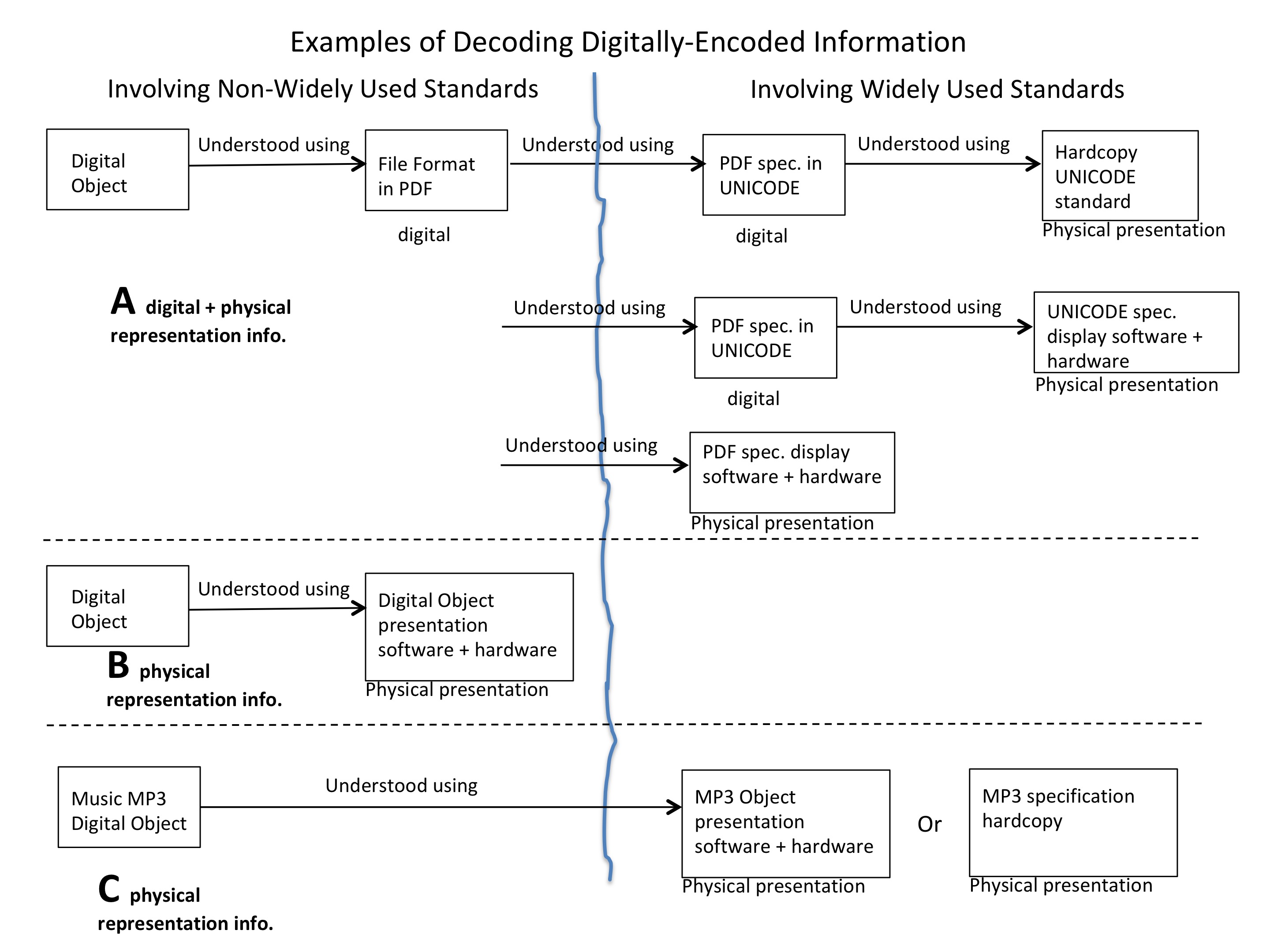


Figure 3-3: Examples of Decoding Digitally-Encoded Information

In section C, the widely used standard is incorporated into the software and hardware that is able to present the information of the MP3 Information Object (music sound). Maintaining preservation for this case is much simpler than for case B because for a widely used standard such as MP3, the format is well known and there will be widely used software and hardware combinations that can be used in transforming the Music digital object into more readily available formats of the future.

This Figure also argues that there are 2 distinct domains of digital Representation information understanding that the Archive needs to consider to accomplish its preservation mission. When a Representation Information component involves the use of a non-widely used standard (or non-standard), Archive staff must be able to present this Representation information to themselves and then understand the presented information in order to be in a position to transform it should this be needed over the Long Term.

When a Representation Information component involves a widely used standard, Archive staff must be able to present this information to themselves in an understandable form but complete understanding can be left to standards experts as long as they are available. This enables the Archive to make transformations over the Long Term without having to be fully knowledgeable of some potentially very complex standards. For such widely used standards, there must by definition be readily available software that incorporates an understanding of the standard and that can be used to present the information, encoded in the associated digital object, to human senses.

For any preserved information that is to be made available to the Consumer, the Archive must also ensure that the information is presentable to the human senses of individuals in a Designated Community. This is essentially arguing that when information has been digitally encoded, it is not enough to simply present the digital object and its decoding information (i.e., Representation Information) as separate entities without having demonstrated the ability to actually perform the decoding of the digital object to present the full information to human senses. While it may be impractical to present the full information for very large Information Objects, at least some information needs full presentation so it cam be compared with separately provided validation Information obtained during Ingest. This presentation ability needs to be maintained and should be a valuable audit criteria. It seems to this author that this should be raised to be an important point and discussion within the OAIS reference model.

For information where the preservation objective is ‘understanding’, the Designated Community must be able to understand the presentation of this information. When this is a requirement, it must be established by the conclusion of the Ingest function. This also means that the Archive has agreed to maintain the understanding in the face of an evolving Designated Community, whatever this may take, unless the preservation objective is changed.

As previously noted, for information where the preservation objective is ‘experience’, the Archive will need to maintain (or ensure access to) the ability to generate the signals (information) expected to elicit that experience. This includes the maintenance of (or access to) necessary hardware and software.

Again, the points above are focused on preservation and not on issues of access and ease of use. These issues will likely drive more migration and especially transformations to keep up with readily available technology, for information presentation, available to the Designated Community.

**3.2 Conclusions**

1. Representation Information is not always an Information Object and the OAIS Reference Information definition and modeling need correcting.
2. All Representation Networks are terminated by the presentation of information to human senses by physical devices.
3. An Archive should be able to present, to itself, all Representation information in a manner perceptible to human senses.
4. An Archive should make an explicit distinction between Representation Information that is a widely used standard (and thus has widely available software for application) and the Representation Information that is not widely used or is unique. This divide allows the Archive to defer full understanding of the Representation Information to external, available, experts. Full understanding is required for non-widely used standards or unique representations. This divide will likely change over the Long Term.
5. An Archive needs to maintain the capability of presenting, to individuals of the Designated Community, the full information content in a form perceptible to human senses. Alternatively, it needs to be able to demonstrate how such individuals can perform this presentation themselves.
6. When preserved information is given the preservation objective of maintaining understanding, it must be understood by a Designated Community. The Archive needs to ensure this requirement is met by the conclusion of the Ingest process.

**4. Role for Information Access and Use**

**4.1 Discussion**

Sections 2 and 3 above are focused on information preservation. Any Archive that has preserved information must also have some concern for access, whether or not it can be determined when such access may be needed. Otherwise there is no point in attempting to preserve information. The current OAIS Reference Model simply requires that the OAIS Archive “make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.” There is no OAIS requirement for any degree of ease of access or ease of use. This is properly the domain of the individual Archive and its stake holders, and can range from ‘no access unless…’, to some specific access and use requirements. The various uses to which accessed information may be put are virtually unlimited and therefore must be a local Archive and Designated Community issue. Even if there is no access planned for the foreseeable future, some type of Designated Community, even if just Archive staff, can be specified.

A given Archive may have been given some specific access requirements. Meeting such requirements could impact how the Archive maintains its preserved information, and changing technology, in use by the Designated Community, may prompt the Archive to perform migrations. Nevertheless, the Archive must continue to meet its preservation objectives as established by the conclusion of the Ingest function, or by a subsequent revision by being able to demonstrate presentations meet ‘experience’, ‘understanding’, or ‘experience and understanding’ criteria as established by the conclusion of the Ingest function, or by a subsequent revision.

**4.2 Conclusions**

* Access and use requirements, to the extent they have been defined, are local to individual Archives and their Designated Communities, and this may influence how information is held and the frequency and nature of migrations.

=======