#### Representation Information

The Representation Information accompanying a digital object, or sequence of bits, is used to provide additional meaning. It typically maps the bits into commonly recognized data types such as character, integer, and real and into groups of these data types. It associates these with higher-level meanings: this includes the description of the, possibly complex, ways objects are interrelated (for example, Representation Information could indicate that three numbers represent temperature, latitude and longitude; and they are expressed in degrees Celsius and angular degrees; and they are interrelated in that the temperature is measured at the specified longitude/latitude).

The Representation Information accompanying a physical object like a moon rock may give additional meaning, as a result of some analysis, to the physically observable attributes of the rock. This information may have been developed over time and the results, if provided, would be part of the Information Object.

There will be special cases where, for a specific Data Object and a specific Designated Community, the Knowledge Base of the Designated Community is adequate for its members to understand or use the Data Object. In such cases the Representation Information could be the statement that no additional Representation Information is needed for that specific Designated Community at this time. Of course the Knowledge Base of the Designated Community may change over time and additional Representation Information may then be required. In order to satisfy the Mandatory Requirements the Archive should ensure that appropriate Representation Information will be available in future, for example details of specific software required or semantics, such as the meaning of spreadsheet columns, which are currently common knowledge. Therefore any Representation Information that can be gathered at ingest should be included since it will likely be more costly to rediscover and add it at a later time.

The remainder of this subsection focuses on the Representation Information object when the Data Object is specialized as a Digital Object.

##### Representation Information Types

The Digital Object, as shown in figure 4‑10, is itself composed of one or more bit sequences. The purpose of the Representation Information object is to convert the bit sequences into more meaningful information. It does this by describing the format, or data structure concepts, which are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc. These common computer data types, aggregations of these data types, and mapping rules which map from the underlying data types to the higher level concepts needed to understand the Digital Object are referred to as the **Structure Information** of the Representation Information object. These structures are commonly identified by name or by relative position within the associated bit sequences. The Structure Information is often referred to as the ‘format’ of the digital object.

The Representation Information provided by the Structure Information is seldom sufficient. Even in the case where the Digital Object is interpreted as a sequence of text characters, and described as such in the Structure Information, the additional information as to which language was being expressed should be provided. This type of additional required information is referred to as the **Semantic Information**. When dealing with scientific data, for example, the information in the Semantic Information can be quite varied and complex. It will include special meanings associated with all the elements of the Structural Information, operations that may be performed on each data type, and their inter-relationships. Figure 4‑11 emphasizes the fact that Representation Information contains both Structure Information and Semantic Information, although in some implementations the distinction is subjective. It is useful to remember that the Semantic Information associated with parts of some digitally encoded information is independent of the format. For example, the meaning of numbers in a data file is independent of whether they are encoded as scaled integers or as IEEE Reals; the meaning of words in a document is independent of whether the document is Word or PDF.

This figure also shows that Representation Information may contain Other Representation Information, i.e. Representation Information which cannot easily be classified as Semantic or Structural. This indicates that the taxonomy of Representation Information presented here is far from complete. For example software, algorithms, encryption, written instructions and many other things may be needed to understand the Content Data Object in ways exemplified by the Preservation Objectives.All of which would by definition be Representation Information, yet would not obviously be either Structure or Semantics. Information defining how the Structure and the Semantic Information relate to each other, or software needed to process a database file could be regarded as Other Representation Information.

Structure Information, Semantic Information and Other Representation Information are both sub-types and components of Representation Information.

One would expect there to be at least Structure Information and Semantic Information, for example English text in a PDF could have the appropriate PDF standard as Structure Information, and the fact that it is written in English as Semantic Information and there could be PDF software as Other Representation Information. Note that the PDF software provides an easier way to use the PDF file.

In general, it is possible for the various pieces of Representation Information to provide alternative ways to add meaning to the Data Object. For example, if the Structure Information and Semantic Information enable the Data Object to be understood, it is possible for the Other Representation Information, such as a piece of software, to provide an simpler way use the Data Object.

Representation Information is an Information Object that may have its own Data Object and its own Representation Information associated with understanding each 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 Information Network**.

As an example, tabular data (numbers and text) in plain text can be formatted as a Comma Separated Value (CSV) file conforming to the IETF RFC-4180 specification. This specification states that the plain text can be provided in ASCII format however the ASCII standard is not provided. To address this situation, the ASCII standard is simply referenced to add the Representation Information that is needed for a full understanding. Therefore the ASCII standard is a part of the Representation Information Network associated with the CSV file and needs to be obtained by the OAIS in some form, or the OAIS needs to track the availability of this standard so that it may take appropriate steps in the future to ensure its CSV file Representation Information is fully understandable.

Figure 4‑11 : Representation Information Object

##### Representation Information Networks

Representation Information, which is itself an Information Object, may be expressed in physical forms (e.g., a paper document) or in digital forms. When the Representation Information is in digital form, additional Representation Information is needed to understand the bits of the Representation Information as described in the previous subsection. In principle, this recursion continues until physical forms, which can be understood by the Designated Community, are encountered. This arrangement of Representation Information is referred to as the Representation Information Network. The Representation Information Network is terminated when the original content object is understandable or usable by the Designated Community which implies that all the Representation Information objects in the Representation Information Network are also understandable or usable by the Designated Community. For example, Representation Information expressed in ASCII needs the additional Representation Information for ASCII, which might be a physical document giving the ASCII standard. Each item of Representation Information can have multiple components, including multiple referenced Representation Information components; each with its own Representation Information.

To preserve the meaning of an Information Object, its Representation Information must also be preserved. This is most easily accomplished when the Representation Information objects are expressed in forms that are easily understandable, such as text descriptions that use widely supported standards such as ASCII characters for electronic versions. One problem with the use of only text descriptions is that such descriptions can be ambiguous. This is addressed by the use of standardized, formal description languages containing well-defined constructs with which to describe data structures. These languages may need to be augmented with text descriptions to convey fully the semantics of the Representation Information.

As the Knowledge Base of the Designated Community changes over time, the Representation Information Network may need to change accordingly. As noted in 2.2, an OAIS has a choice of whether to collect all the relevant Representation Information or to reference its existence in another trusted or partner OAIS Archive; this is an implementation and organization decision.

The Content Information must be defined and separated into Content Data Object and Representation Information. It is again an implementation and organization decision related to the way Data Objects are ingested and stored in the OAIS. The OAIS may define a single Content Data Object as a collection of separate Data Objects or those individual Data Objects could each be considered to be a Content Data Object. The Representation Information required will clearly be related in these two cases and would have to describe the relationship between the Data Objects.

For example, in the case of performing arts, the Content Data Object may be the score as a PDF document, and the Representation Information would include whatever information is needed to re-perform (as the way to use and understand) the piece, such as the way to display the PDF file, the audio processing software needed, placements of hardware such as loudspeakers, movement directions, and a description of how these relate to each other and to the Content Data Object, each of which may be quite complex, encoded in a separate way, and not easily described either simply as Structure or as Semantics. Alternatively, the Content Data Object may be multiple Data Objects including the score, the audio processing software needed, placements of hardware and movement directions. Each of these Data Objects will have its own Representation Information and there will need to be additional Representation Information that describes how the several Data Objects are related.

Two special types of Representation Information are **Representation Information Rendering Software** and **Access Software.** Representation Information Rendering Software is able to display the Representation Information in understandable forms. For example, the file and directory structure of many CD-ROMs conforms to ISO 9660. This standard is Representation Information describing how these CD-ROM file structures are implemented, and it may be obtained as a paper document. However, it may also be obtained as a digital object that needs to be understood as a PDF object. Rather than actually obtaining the documentation of PDF and writing software to understand the ISO 9660 object, an OAIS may use available PDF display software to render the ISO 9660 documentation humanly visible and readable. In this role the PDF display software is referred to as Representation Information Rendering Software because it is used to render the Representation Information. It may also terminate the Representation Information Network, because it is sufficient to display all the Representation Information needed to understand or use the Data Object. If the OAIS does not also obtain the associated description of PDF, it needs to record and track this fact because when PDF objects are no longer cost-effective for access and display, the ISO 9660 documentation expressed as a PDF object will need to be migrated to a new form.

Access Software presents some or all of the information content of an Information Object in forms understandable to humans or systems. It may also provide some types of access services, such as displaying, manipulating, processing, or sub-setting, to an Information Object. For some types of Digital Objects, such software may be widely available. It is not necessary for the OAIS to maintain or provide such software. The OAIS may want to maintain and provide this software for more specialized types of Digital Objects.

Since Access software will incorporate some understanding of the Representation Information, some Archives may attempt to use Access Software as a substitute for full Representation Information. Access Software source code, which embodies at least a partial understanding of the associated Representation Information, may be used as documentation expressing such Representation Information. A problem with this approach is that the desired Representation Information may not be clearly identifiable as it may be mixed with various processing and display algorithms, and may be incomplete since the code assumes an underlying operating environment. It may be difficult to tell, from the software code, what Representation Information is missing. The use of Access Software executables, without the source code, such as may occur with proprietary formats, presents a much greater risk for loss of information because it is more difficult to maintain an operating environment for software than to migrate documentation over time. The practical use of emulation techniques to preserve working software is an area of active research. This is a significant issue for those desiring to preserve a look and feel to information access. Migration and software preservation are discussed more fully in section 5.