An Analysis of Information Migration

DMS 98-01-25

The purpose of this paper is to provide a more detailed assessment of the migration problem and its relation to OAIS data modeling concepts. It also addresses a key concern from the Panel 2 workshop which was to provide scenarios identifying the key reasons for migration. It is hoped that from this paper, and comments to it, we can rapidly update the OAIS migration section 5. Missing from this paper is the issue of Preservation Description Information update during migration - when is it needed and when is it not needed. Also, some additional discussion on how to check for valid logical Repackaging could be helpful.

We define information migration as the movement of digital information, while intending to preserve it, within a given hardware and software environment that is under the control of a single administration. This administration has control of all hardware and software configuration issues within this environment. A single OAIS provides such an environment. This environment may be characterized, from a migration perspective, with a model as given in section 1.

1. Migration Environment Model

The purpose of this section is to develop a simple model of the hardware and software environment in which migration is likely to take place so that key terms and concepts can be understood. An important issue is the extent to which this model is adequate, even when it is restricted to a view of file management systems, both simple and sophisticated. Does this model work for something like Open Vault?

Digital information resides on physical media within the environment. It is assumed to be written to the media by a writing device controlled by driver software, and it is read by driver software controlling a reading device. The reading and writing devices may, or may not, be the same device. .

File management software, working with the driver software, writes and/or reads a particular media format. This format is independent of any application data that will be stored on the media. Examples of such media formats include end-of-files written on unlabeled 9-track tape, ANSI standard labels written on 9-track tape, ISO-9660 volume/directory structures on CD-ROMs, mass storage vendor specific formats on DLTs, and VMS file formats on tape. As a short hand, we can refer to all these formats as being written by corresponding file-management systems. Each file-management system provides a set of services to applications that want to store information on the media. The information to be stored on the media in the media format is provided as a set of attributes. There is a primary attribute, which is the content of a file as a byte-sequence. There are also supporting attributes, both for files and for media volumes.

At a minimum, all the file-management systems are assumed to provide the ability to store and retrieve multiple sets of application-provided byte sequences, which are stored as primary attributes. Most of the file-management systems also provide a range of related

services which accept additional application provided information and return it upon request. These are the supporting attributes. Examples are volume names, directory names, file names, file content attributes, creation dates, and access permissions. These additional services, and their supporting attributes, are not standardized across media types, and often vary across vendor file-management systems dealing with the same media type. This poses a challenge for information migration across these file-management systems.

To examine migration issues more closely, we will address the migration of Content Information within an OAIS. The Content Information is defined by Representation Information that applies to a set of bits. This set of bits resides on one or more media volumes. Access to these bits requires the use of the Packaging Information, which may be logical or physical or both. Physical Packaging Information is the media volume(s) holding the Content Information. Logical Packaging Information is that set of bits, typically on the same media as the Content Information bits, which delimits the Content Information bits and thus supports access to the Content Information. For example, if the Content Information is defined to be the content of a file, then the immediate Packaging Information bits are those that implement the file structure on the medium. This Packaging Information is a part of the media format.

How far the Packaging Information is considered to extend, such as including all of this media format and the underlying media volume, is a local decision. In practice, it will likely be defined to be the entire media format because this is what is usually needed by the access software. In the most general case, some of the logical Packaging Information may be on media other than those used to hold the Content Information, if needed to fully delimit the Content Information. For example, information needed to tie two parts of a Content Information object together, when those two parts are on separate media volumes, may reside on a third volume.

The boundary between Content and Packaging Information, as instantiated in specific cases, plays a critical role in establishing cost-effective migration strategies as will be described below.

2. Migration Drivers

In an OAIS there are several drivers that can cause the primary copy of a Content Information object, or part of that object if it is spread across multiple media, to be migrated by reading it from media, possibly modifying it while minimizing information loss, and then writing it back to media and making this new version the new primary copy. These include:

- **Media Replacement**: The need to replace the original media, whether fixed or removable, with new media (and possibly devices) to improve media characteristics such as reliability and performance, or to reduce costs. The old media, and possibly devices, are removed from the environment. Examples include replacing older 9-track magnetic tapes with newer 9-track tapes to reduce the probability of tape read errors; replacing 9-track drives and tapes with a robotic system handling 3480 cartridges to

reduce human intervention; or replacing CD-ROMs with DVDs to obtain higher capacity on a single media volume.

- **Device Usage**: The need to move the Content Information to different devices, and to their associated media and media formats, to obtain different performance or cost situations. A prime example is the movement of Content Information from a hard disk to magnetic tape to reduce storage costs, or in the reverse direction to gain access performance.

- **New Management System**: The need to use a more cost-effective, or more widely supported, file management system when the new system requires a different media format from that used by the old system. For example, management system A uses a proprietary media format for the DLTs that it manages, while new management system B uses a different media format in managing the same devices. Another example is VMS management of VMS-attributed tapes versus UNIX management of UNIX-attributed tapes.

- **New Packaging Formats**: The need to maintain the Packaging Information in a different format which is more cost-effective for migration, or for the eventual users of the information, while simultaneously ensuring that the required packaged-content (e.g., Content Information) inter-relationships are not lost. For example, a file system can be used to package file content with a file name. To remove file system dependencies and improve migration ease, a generic packaging structure, which can reside completely in a file's content, is defined and it associates a name with a byte-sequence. The migration to new packaging formats, either physical or logical, will be referred to as **Repackaging**.

- **New Packaged-Content Formats**: The need to maintain the Content Information (or the Preservation Description Information) in a different format which is more cost-effective for migration, or for the eventual users of the information, while simultaneously minimizing information loss. This is referred to as **Transmutation**. Examples include
    - replacement of EBCDIC codes with ASCII,
    - replacement of IBM 7094 floating point with IEEE standard floating point,
    - replacement of an older science data format with new science data format,
    - replacement of an older word-processing document format with a new word-processing document format, and
    - loss-less compression of the Content Information bits.

The impacts of these drivers depend critically on the definition of the Content Information and the way in which it is distributed among the attributes of the file-management system. It is recognized that these drivers are not independent. The following section takes a structured approach to assessing their impacts and to clarifying the interdependencies.

3. Selected Information Migration Cases

There are several information allocation and handling approaches that, together with the migration drivers, can be used to identify a range of migration scenarios for preserving Content Information. The approaches adopted are:

- Media Format Preservation,
- Supporting Attribute Fixed-Content Preservation,
- Primary Attribute Fixed-Content Preservation,
- Content Information Reformatting (i.e., Transmutation).

Each of these will be defined and examined as to the impacts of the various migration drivers on the migration of the Content Information. Other combinations should be readily derivable from these.

3.1 Media Format Preservation

In this approach, the Content Information has been defined to include the media format and all bits contained therein. This could arise if the media, and all its content, was historically important. For this Content Information, the Packaging Information is the media volume itself. The impacts of the migration drivers are as follows:

Media Replacement

a.) Movement of this Content Information to the same media type is generally straight forward as long as the media type is still supported by vendors. A copy operation will generally be supported by the file-management system. The copy is accomplished by reading and writing some of the information and regenerating other information that is known to be standard for that particular media format. This may result in slight differences, at the bit level, of some standard media format attribute values between the original and the copy, which could happen if the media format included such attributes as a manufacture sequence number or a 'first date of writing' field. In any event, the resulting media is directly substitutable for the original to the file-management system, and the media format, with its content, has been effectively preserved. We can refer to this type of migration as **Information Replication** (IR). It is a special case of a more general migration approach called **Repackaging** where the Packing Information (in this case, the media volume) changes.

If the file-management copy does not provide total media format equivalence (bit-for-bit), and this is strictly required, then it will be necessary to develop special techniques to read and write all the media format bits. This may require writing special driver software. It may be very difficult to access all the bits as some may be written and read by hardware or firmware and not passed to the driver.

b.) Movement of this Content Information to different media types is possible if access to

all the media format bits can be achieved. The entire bit-sequence can then be copied into the primary attribute of a file on another media type. This type of Repackaging can be called **Media Format Wrapping**. Direct access to the original media format content would require special software. Alternatively, it may be copied back to the original media type if that type and format is still supported by vendors.

Device Usage

Preserving the media format while migrating it to a new device and its new media format can be accomplished using the Media Format Wrapping approach when all the original media format bits can be read and copied. This would not provide improved access, but it could be used to reduce storage costs. If all the media format bits can not be accessed, migration is not possible.

New Management System

If a new file-management system supports the device and the media format that is to be preserved, then no migration takes place. Otherwise the Content Information must be migrated as described under Device Usage immediately above.

New Packaging Format

Since the media format is preserved, the original packaging information is the physical media itself. Migration to the same media type provides Repackaging with a new physical media of the same type. This can be called physical Repackaging.

Migration to a new media type, when it can be accomplished, is a case of Media Format Wrapping and it results in new Packaging Information on the new media. This Packaging Information consists of the bits used to wrap, or encapsulate, the media format bits. It can be extended to include all the media format bits of the new medium, and also the new physical medium itself, if desired.

New Packaged-Content Format

By definition, this media format preservation does not allow Transmutation.

3.2 Supporting Attribute Fixed-Content Preservation

In this approach, the Content Information has been defined to include some or all of the values in supporting attributes, as well as the primary attributes. This means that the media format does not need to be preserved, but some of the values it contains do need preservation. In addition, the representation of these values is considered fixed. This is a very common case and a prime example is the use of a file name, which is a supporting attribute, to hold important information (deemed to be Content Information) about the file content. The impacts of the migration drivers are as follows:

Media Replacement

a.) Movement of this Content Information to the same media type is straight forward as long as the media type is still supported by vendors. A copy operation will move the content of supporting attributes, and the primary attributes, without alteration and preservation will be achieved in the migration. This is Information Replication and is a form of physical Repackaging.

b.) Movement of this Content Information to different media types requires that those values of supporting attributes and primary attributes, defined to be Content Information, have the same representations before and after the movement. For example, if the supporting attribute 'file name' for some primary attribute on some medium is defined to be part of the Content Information, and it is implemented as an ASCII string, then this ASCII string can be moved into a new overall format as long as its relation to the primary attribute's content is maintained - i.e., it is still a supporting attribute. This is a form of logical Repackaging.

Device Usage

Moving the supporting attributes and primary attributes to a new media format to take advantage of device characteristics has the same impacts as the replacement of media with a new media type, as described immediately above.

New Management System

If a new file-management system supports the device and the media format, then no migration takes place. Otherwise the Content Information must be migrated to a new media type as described above under Media Replacement.

New Packaging Format

a.) Migration to the same media type provides Repackaging with a new physical media of the same type. This is a case of physical Repackaging.

b.) Migration to a new media type, which only has to preserve the content of some supporting attributes and the content of the primary attributes, results in a new set of packaging bits for this content. The mapping of this content into the new media format can take many forms, and each can be considered a valid packaging object as long as the required Content Information inter-relationships are preserved. These are generally types-of-associations among pieces of the Content Information, such as the relationship of names to objects.

New Packaged-Content Format

By definition, this supporting attribute preservation does not allow Transmutation.

3.3 Primary Attribute Fixed-Content Preservation

In this approach, the Content Information has been defined to reside in the content of primary attributes. This means that the media format and the content of supporting attributes do not need to be preserved, and only the values of primary attributes need preservation. In addition, the representation of these values is considered fixed. This is a very cost-effective case for migration because it is not necessary to preserve such supporting attributes as file names or directory names. If such names were needed as part of the Content Information, they have already been included appropriately in the content of primary attributes. The impacts of the migration drivers are as follows:

Media Replacement

a.) Movement of this Content Information to the same media type is straight forward as long as the media type is still supported by vendors. A copy operation will move the content of the primary attributes without alteration and preservation will be achieved in the migration. This is Information Replication and is a form of physical Repackaging.

b.) Movement of this Content Information to different media types requires that those values in the primary attributes, defined to be Content Information, have the same representations before and after the movement. Since all file-management systems support a sequential copy into the primary attribute, this migration is straight forward. This is a form of logical Repackaging.

Device Usage

Moving the primary attributes to a new media format to take advantage of device characteristics has the same minimal impacts as the replacement of media with a new media type, as described immediately above.

New Management System

If a new file-management system supports the device and the media format, then no migration takes place. Otherwise the Content Information must be migrated to a new media type as described above under Media Replacement.

New Packaging Format

a.) Migration to the same media type provides Repackaging with a new physical medium of the same type. This is a case of physical Repackaging.

b.) Migration to a new media type, which only has to preserve the content of the primary attributes, results in a new set of packaging bits for this content. The mapping of this content into the new media format can take many forms, and each can be considered a valid packaging object as long as the required Content Information inter-relationships are preserved. A common repackaging example is moving the byte sequence from the content

of one file on a medium to the content of another file on another medium. The new packaging consists, at a minimum, of the file structure bits on the new medium. This is logical Repackaging.

New Packaged-Content Format

By definition, this primary attribute preservation does not allow Transmutation.

3.4 Content Information Reformatting (i.e., Transmutation)

In this approach, the Content Information has been defined to allow changes to some or all of the representations used. This differs significantly from the three previous approaches which all required representations to be unaltered during migration.

Each discrete representation can be modeled as consisting of a base set of entities, a set of resulting entities, and mappings between the entities. For example, the ASCII representation has a set of 7-bit sequences as its base entities, and a set of characters as its resulting entities, and mappings that relate 7-bit sequences to characters. More complex representations, involving groups of primitive data types and their inter-relationships, that result in complex formats, can be put into the same model of base entities, mappings, and resulting entities.

For some Content Information, information is adequately preserved even when some of the base entities and mappings in some representations are altered, as long as the resulting entities are adequately retained. The adequacy of retained entities is often subjective and whether it is acceptable to allow a new representation depends on the specific case. We refer to the alteration of representations, when it applies to Content Information or Preservation Description Information, as Transmutation.

There are two types of Transmutation that can be seen: Strict Transmutation and Loose Transmutation. **Strict Transmutation** occurs when a representation defines a resulting entity that is strictly semantically-equivalent to a resulting entity defined by a different representation. An example is replacing the ASCII code for “A” with the UNICODE code for “A”. **Loose Transmutation** occurs when Strict Transmutation can not be guaranteed. Replacing an IBM 7094 floating point value with an IEEE floating point value is one example because the entities they represent are not strictly equivalent. One will have more precision than the other. However they may be sufficiently equivalent, depending on what the values they represent are being used for, to be effectively interchangeable. If this is the case, Loose Transmutation effectively preserves the information content. For complex formats, where the meanings and relationships among groups are significant, it may be difficult to establish that Loose Transmutation has adequately preserved Content Information. The impacts of the migration drivers are as follows:

Media Replacement

a.) Movement of Content Information, regardless of where it is stored in the media format,

to the same media type is straight forward as long as the media type is still supported by vendors. A copy operation will preserve the media format and move the content of all attributes without alteration. There is no need to use Transmutation.

b.) Movement of Content Information to different media types can be simplified when the representations of some information items can be changed. For example, consider a CD-ROM where it has been determined that the Content Information consists of the byte-sequences in each file together with the names of each file and each directory, and their associations. Note that the representations of the file and directory names do not need to be preserved, nor does the implementation of the file and directory structures need to be preserved (i.e., no need to preserve any of the CD-ROM media format). This Content Information can be readily migrated to a DVD medium because its media format supports the name strings and associations, and it will hold the byte sequences in its files. Other directories and files could also be added to the DVD because of it larger capacity. In either event, there has been a Repackaging of the Content Information. There may also have been a Transmutation since no attempt was made to ensure that the representations of the file and directory names were the same between the CD-ROM and the DVD.

c.) As a variation on this example, the Content Information could also have been moved into a single, large, file on another medium. The format of this file would have to be organized to recognize the byte-sequences from each CD-ROM file, the file and directory names, and their inter-relationships. Any representation, including UNICODE, could be used for the directory and file names. The structure of this file provides the Packaging Information for the Content Information, and except possibly for the size, provides an easily migrateable form. This is another example of Repackaging, possibly including Transmutation.

Device Usage

Moving Content Information to a new media format to take advantage of device characteristics has the same impacts as the replacement of media with a new media type, as described immediately above.

New Management System

If a new device-management system supports the device and the media format, then no migration takes place. Otherwise the Content Information must be migrated to a new media type as described above under Media Replacement.

New Packaging Format

The ability to incorporate representation changes to some of the Content Information can assist in migrating to new Packaging Formats, as described above under Media Replacement.

New Packaged-Content Format

This is not applicable because this is the same as the approach being considered and is redundant.

4. Summary and Conclusions and Issues

This analysis has presented a model of the migration environment, including the relationships of Content Information and Packaging Information. It has attempted to identify key migration drivers, and several ways in which Content Information may be specified in the migration environment. It has looked at the interactions of the migration drivers with the ways of specifying Content Information to determine possible approaches to migration. We can draw the following conclusions:

o The fundamental and useful migration concepts, which fits with our data modeling terms and concepts, are Repackaging of Packaging Information and Transmutation of Content Information (or Preservation Description Information).

o The distinction between Packaging Information and Content Information, when the Content Information is not all contiguous, requires that the Packaging Information provide some associations among the split pieces. These associations must be retained in any Repackaging, unless they are instead added to the Representation Information for the Content Information through an allowed Transmutation. In any event, the Packaging Information should have only the simplest associations (e.g., this is an attribute of that), leaving the complex association semantics to the Representation Information. This facilitates Repackaging.

o When possible, Content Information definitions should make clear which representations are potentially replaceable because the base entities are not fundamental for information preservation.

o From a migration perspective, in a file management environment, all Content Information (and, by inference, Preservation Description Information) should be stored in the content of file byte-sequences, and not in supporting attributes. The penalty will be reduced access performance.

In support of this conclusion, it is noted that the supporting attributes vary by media type partly in response to media and device characteristics. However for long term information preservation, where the effort to migrate may be a significant cost factor, there is a strong driver to be able to migrate information across media types in an automated manner. Current mass storage management systems typically support ‘device usage’ migration for a limited set of devices. However they also usually use proprietary formats that result in the need to migrate all the stored information to new media formats when the management system is replaced. This is costly and may be impractical for very large stores. When this occurs, someone needs to define a mapping from the old media format to the new media format, including supporting attributes. If some of the information to be preserved has been encoded into supporting attributes, and is not also included in the ‘file content’, much

greater care has to be taken with the mapping to ensure information is not lost. On the other hand, if all the information to be preserved is within the 'file content', then the mapping is only concerned with performance issues and not with information preservation as well.

Issue: The basic model of the migration environment assumes it varies across simple to sophisticated file management systems. It does not address migration of content in a DMBS. What should we say about this in the reference model? Is there an easy extension to DBMS (probably not).

Issue: Preservation Description Information update during migration - when is it needed and when is it not needed?