A Collaboration Model between Archival Systems to Enhance the Reliability of Preservation by an Enclose-and-Deposit Method

Koichi Tabata, Takeshi Okada, Mitsuharu Nagamori, Tetsuo Sakaguchi, and Shigeo Sugimoto

Graduate School of Library, Information and Media Studies, University of Tsukuba, 1-2 Kasuga, Tsukuba, 305-8550 Japan

{tabata, okadatak, nagamori, saka, sugimoto}@slis.tsukuba.ac.jp

Abstract. This paper proposes a model of cooperating archival systems that mutually deposit archived resources in order to enhance the reliability of long-term preservation. In the model, a resource and its preservation metadata in a source archive are enclosed together in a bag, and the bag is deposited in a destination archive as a resource. The proposed model is a simple approach, but it is advantageous for the following reasons: (1) Duplication of resources enclosed in a bag preserved under different environments enhances the reliability of archiving; (2) Simplicity is an essential factor in making the technology adoptable in small archives and in making preservation more reliable; (3) No archival systems are permanent, especially those created by small organizations; (4) Re-writing or transcribing metadata is not necessary. This paper first discusses the requirements for collaborating distributed archives and then proposes a simple "enclose-and-deposit" model. The model is examined for both deposit of a single resource and deposit of a collection of resources. The last section describes an implementation of the model, using XML as the encoding scheme and DSpace as the underlying architecture of the collaborating archival systems.

1. Introduction

Precious old Japanese material that no longer exists in Japan has often been reported to have been "discovered" in an archive in the United States, implying an effective method for the long-term preservation of information resources. Analysis of the specific circumstances often reveals that an archive in Japan did not try to preserve that material, not realizing its importance for the future, or lost the material despite the intention to preserve it. In the American archive, however, it is assumed that such material is important even though its contents are unclear. The person in charge at that time, despite his/her inability to understand Japanese, typically encloses the material in a storage bag and preserves the bag somewhere in the archive after noting on its surface, in English, the origin of those materials. Re-discovered later, this material's importance is recognized anew; it is formally registered as belonging to that archive; and an explanation of the contents is provided in English. Such material may be very important for Japan as well, therefore, a decision is often made to store a copy of the material in a Japanese archive also.

Such a process occurs even if the archived material is a digital resource. Many resources are created and published on the Internet. Deposit libraries are making efforts to collect and archive these resources, but it is difficult to evaluate their importance and to describe detailed metadata as the OAIS reference model recommends. The only organization (or person) that can appropriately describe the metadata required for preservation of a resource is the one that has created the resource. However, it is difficult for the organization to run its own archival system in a secure environment and for a very long time, except in such cases as legal deposit libraries and large-scale governmental institutions. Collaboration among those organizations is a key to realizing reliable archival systems.

We borrow the phrase "Lots of Copies Keep Stuff Safe" from the LOCKSS project [1] to encourage the creation of multiple legal copies of archived resources in order to enhance the reliability of archives. The OAIS reference model does offer a highly reliable self-contained model to preserve digital

resources in a single system. However, in the Internet environment, many organizations run their own archival systems using open source software such as DSpace, and not all archival systems are as reliable as they should be. Thus, in order to enhance the reliability of preservation, we need to build a network of collaborating archival systems that mutually deposit copies of their valuable resources.

2. Requirements Analysis for Collaborating Archival Systems

This section analyzes the scenario presented in the previous section in order to establish requirements for a model of long-term preservation by collaborating archives.

(1) An archive in Japan plans long-term preservation of their valuable digital resources. However, if a resource is preserved in only the local archive, it may be lost someday, so they decide to deposit copies of important items in another archive in the United States. Because the destination archive for depositing the resources also has the responsibility of preserving digital resources, the survival of the resources deposited from Japan is highly probable.

(2) If that resource in the local archive is lost later, it can be recovered from the destination archive.

(3) Still later, not only the name of the resource but also the name of the destination archive is no longer in people's memory. At some point, the resource will happen to be found in the destination archive, and the resource will be restored to the source archive. If the source archive no longer exists, the destination archive may preserve the resource and offer it for public perusal, or transfer it to an equivalent archive in Japan.

With regard to such a situation, it is important to investigate several issues for preservation by collaborating archival systems.

(a) The resources deposited into the destination archive should not be directly provided to end-users by the destination archive, because the destination archive is a back-up function for the source and the audience community of the destination may not overlap that of the source archive. Thus, the destination archive is not a mirroring archive of the source.

(b) In general, the metadata schemata of the archives will differ. In this example, the languages used to describe the metadata are different. To handle the deposited resource in a schema equivalent to that of its own archive, transcription and/or translation of the metadata from the original form into the target schema must be performed to transfer each deposited resource. However, the costs for transcription and translation should be avoided unless the destination archive directly provides deposited resources for the end-users.

(c) Regardless of the contents of the resource or the metadata from the source archive, the destination archive will receive the deposited resource as well as the metadata from the source archive (the Japanese archive in the above case), all enclosed in a bag, and will add the note "Deposited Resource P" where P is a simple identification number. Also noted on the surface of the bag will be a description of when and from what archive the resource was received, using the description language (English in this case) and the metadata system of its own archive. For example, the added notes may read [Title: Deposited resource P], [Creator: Source archive]. If the person in charge does not understand the metadata system or the language (Japanese) of the source archive, the metadata (in English) for its own archive is prepared for the document describing the method (in Japanese) for opening the bag, and is stored with the bag.

(d) When the destination archive receives a request to return a resource, the deposited resource will be located by referring to the metadata of its own archive and returned as enclosed in the bag, together with the document indicating the method of opening. The source archive will open the bag by referring to the document, take the resource and metadata from the bag, and store them in its own archive.

(e) In the future, if the source archive no longer exists, the corresponding bag will be identified and opened at the destination archive to investigate the resource and its metadata. In the previous example, it is natural to assume that the destination organization could hire a person who understands

Japanese or could request assistance from another archive that has experiences in handling Japanese resources. The resource is then registered formally as an original resource of its own archive, depending on the necessity of offering it for public perusal. The metadata in Japanese will be rewritten into the proper system for its own archive, and the description language will be translated into English. Because sufficient time will likely have passed since the resource was created, there should be no copyright problem; therefore, the resource can be offered for public perusal.

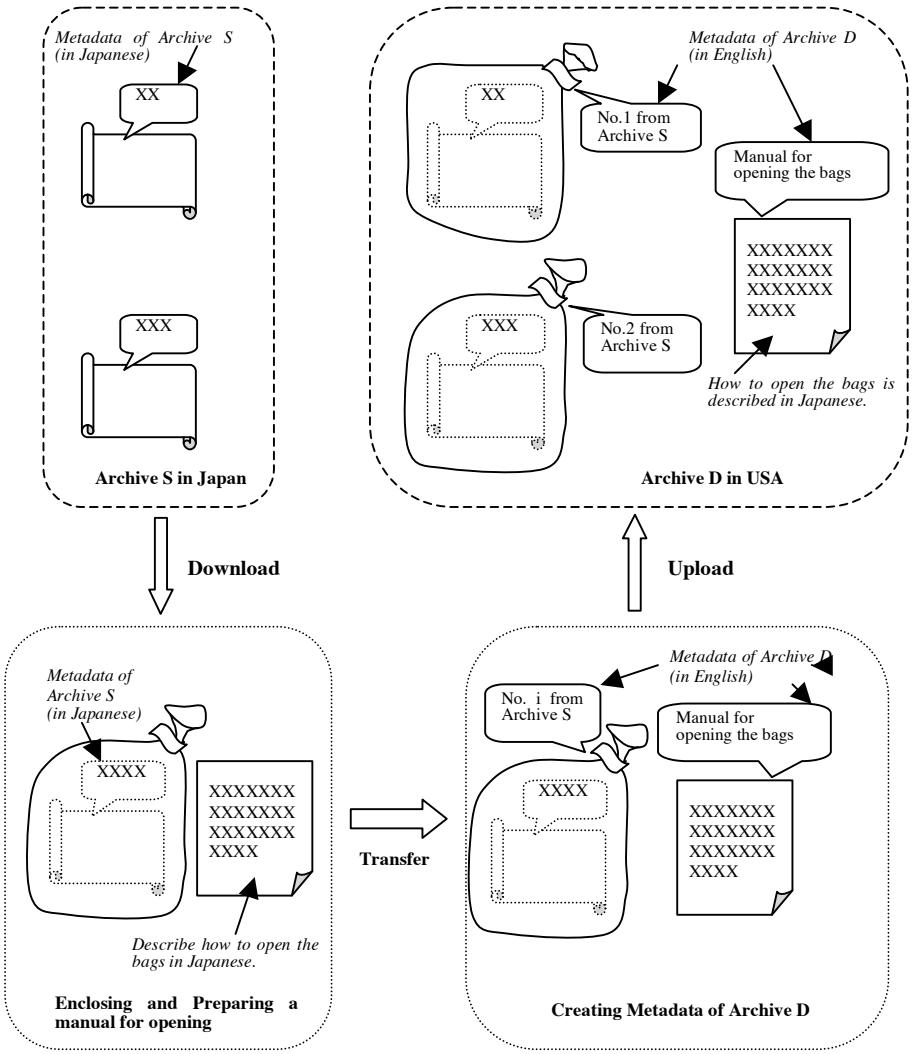

Fig. 1. Basic Concept of Depositing between Archives: On the source side, a resource and its preservation metadata in Archive S are enclosed together in a bag, and a prepared manual describes how to open the bag. The bag and the manual are transferred to the destination, and deposited in Archive D with D's preservation metadata.

This model is presented in Fig. 1. Archive S, the source archive in Japan, selects the resource and metadata to be deposited. Enclosing both of them in a bag and preparing a manual for opening the bag, the source archive sends them to Archive D in the United States. Archive D preserves the bag and its manual for opening by attaching the proper metadata. The metadata to be attached to the bag may be simple, indicating only the source archive (e.g., Archive S, the date of deposit, and the serial number or identifier of the item). The metadata for the manual for opening should describe the method for opening the deposited resource. The return, disclosure, and opening will be performed at a later time by referring to the manual.

This scenario emphasizes the differences between archives in different countries (e.g., those in Japan and those in the United States). However, even in the same country or in regions where the same

language is spoken, differences exist between archives that serve different communities. As mentioned in OAIS, their metadata schemas differ, depending on the properties of the data involved. For example, the data properties for the archive of a university library and those for that of a space observation center are not the same; however, they can collaborate to preserve resources.

Requirements determined from the above scenario are as follows.

(i) Neither the source archive nor the destination archive is required to add any special functions beyond those vital to the management functions of the archive. No functions other than download and upload will be used.

(ii) The work environment to transfer the archival information packages between the archives does not have to be retained and would be set up only when required. Even long afterwards, the work environment for deposit and return work may be re-established by referring to the manual for opening, which should be retained with the resource. Thus, the only items that have to be preserved are the bag and its manual for opening. As long as these items are maintained by the destination archive, the resource can be recovered.

(iii) The metadata schema and the description language may differ between the collaborating archives, but mutually converting their schemas is unnecessary. Deposit can easily be accomplished even between OAISs that preserve different forms of data.

(iv) The deposited resources can be re-deposited to a third archive. The bag can be placed in a new bag and deposited in the third archive. The metadata attached to the bag will describe the deposit history and the required function to recover the enclosed resource. This feature clearly delineates the difference between this model and the conventional backup functions, including mirroring.

3. Collaborative Inter-Archive Deposit of Archival Information Package

In the ISO reference model for an Open Archival Information System (OAIS), an OAIS is defined as an archive that has accepted the responsibility for preserving information and making it available for a Designated Community [2]. These communities are diverse, and the properties of the information and data stored by individual communities differ from each other. Nevertheless, the OAIS reference model specifies a framework of guidelines to be considered as the minimum requirements from the viewpoint of long-term preservation, while allowing individual communities to establish free archival forms outside the framework. In addition, the OAIS reference model presents several models for consumer services provided through the mutual linkages of different archive systems. However, it has not yet presented a model representing long-term preservation through the mutual linkages of these systems.

In this paper, we propose a model projecting additional long-term preservation through mutual linkages between archive systems that comply with the OAIS reference model, based on the discussions described above. Once an archive system has been established, the community will endeavor to maintain it indefinitely by involving the whole organization, so long-term archival resources must be sufficiently protected in order to survive. If such archive systems can mutually deposit particularly important resources, the resources will be more reliably preserved. However, metadata schemas differ from one community to another, so rewriting the metadata of resources deposited from a source archive into the destination archive will require a tremendous volume of work. In this paper, we propose a system to enable the mutual storage of resources between different archives without necessitating such work.

In the OAIS reference model, information to be preserved is expressed by means of an Archival Information Package (AIP) (Fig. 2). An AIP is a conceptual container holding two types of information: Content Information (CI) and Preservation Description Information (PDI). A CI consists of the Content Data Object (CDO) and its associated Representation Information (RI) needed to make the CDO understandable to the Designated Community [2].

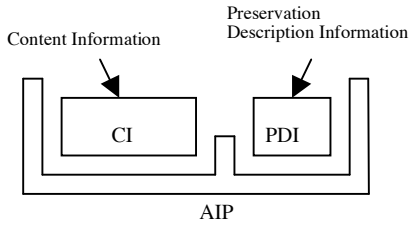

Fig. 2. Archival Information Package (AIP)

The resource and its metadata for preservation correspond to the CI and PDI of the AIP, respectively. The packaging of the resource with its associated metadata in a bag by Archive S and the preservation of the bag by attaching the metadata for Archive D by Archive D correspond to the following process (Fig. 3).

- The source archive (OAIS S) encloses the CI and the PDI in a bag,
- the destination archive (OAIS D) puts the bag in the CI of its own AIP.
- The PDI at the destination is not rewritten from that at the source but contains descriptions such as "Item P deposited by OAIS S on XX YY ZZ" (P is the identification number).
- The position where the bag is put at the destination is, to be precise, the CDO section of the CI.

The manual for opening prepared by Archive S corresponds to the following process; the manual for opening prepared at the source is put in the CI of the AIP of the destination system, and the metadata created to enable the destination system to identify it as the information for opening is put in the PDI. As in the above case, to be precise, the manual is put in the CDO section of the CI.

This nested architecture help avoid complicated tasks to interoperate the source and destination archives and allow "on-demand disclosure" of the packaged information by the destination archive.

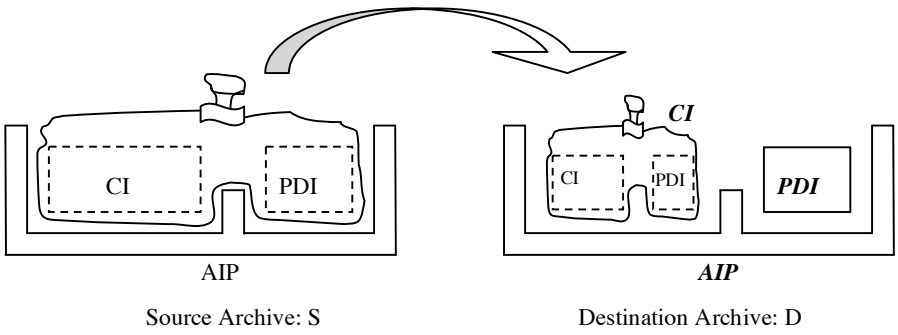

Fig. 3. Concept of Deposition between OAISs

4. Model of Inter-archive Deposit of a Single Resource

Before describing the model in more detail, we define the terms listed below.

(a) "Enclosing" means to express the contents of the AIP of the source archive, i.e. the contents of its components, CI and PDI, as a single XML document. This document is called the enclosing document. Before being enclosed in the XML document, the bit-string portions are converted into character strings by base64Binary. The character set used in the document is specified by the XML encoding declaration.

(b) "Opening" means to extract the contents of the enclosed AIP, i.e. the contents of its components, CI and PDI, from the enclosing document. Character strings converted by base64Binary are reconverted into the corresponding bit strings.

(c) "Manual for opening" refers to the document that enables opening the enclosing document by describing how the enclosing document was prepared.

(d) “Disclosure” means to view the contents of the deposited enclosing document on the destination archive. The enclosing document is an XML document, so it is possible to view its contents by preparing the character set environment as specified by the XML encoding declaration.

Let us now explain the method for preparing enclosing documents in more detail. An enclosing document is prepared by expressing the contents of the AIP of the source system, i.e. the contents of its components, the CI and the PDI, with a single XML document as shown in Expression (1). In the CDO section, the character string being converted using base64Binary is enclosed by <base64Binary> and < /base64Binary >.

```
<?xml version=”1.0” encoding=”xxx”?>
<AIP>
  <CI>
    <CDO>xxx</CDO>
    <RI>xxx</RI>
  </CI>
  <PDI>xxx</PDI>
</AIP>                                                    (1)
```

Expression (1) represents a skeleton of an XML instance of an enclosing document. The <CI> contains one or multiple data objects of different formats. In the case of multiple data objects, the element of the <CI> is expressed <subCI>xxx</subCI>…<subCI>xxx</subCI>. Example of multiple data objects is a hypertext document which is composed of a single HTML text and one or more JPEG images. Sometimes the RI element may be omitted, for example, in the case that the file name of a resource has a file extension, such as ‘abcde.html’ and ‘pqrst.jpg’.

Expression (1) shows a general form of an AIP encoded in XML for transmission between the archives. The schema to encode a resource into an XML instance depends on each archive system. A document which explains the schema of the XML instance must be prepared by the source archive. The document will be referred to by the source archive to open the enclosing document and restore the resource. This document is called the “manual for opening”, which basically consists of character text, and also contains a conversion method, e.g. base64Binary. The destination archive need not interpret the document unless it is explicitly requested to disclose an AIP and to reformulate the resource.

Deposit work - The destination archive receives an enclosing document from the source archive and creates a CI which is composed of the CDO sent from the source as an XML instance and an RI which states that the CDO is an XML instance. Then, the destination archive creates an AIP by packaging the CI with a description about the CI for preservation in the destination archive, i.e. PDI. Description in PDI depends on the policy of the destination archive but a simple example would be a text which tells the fact that the destination archive received the resource from another archive, e.g. “Item P deposited by OAIS S on XX YY ZZ” (P is the identification number).

A manual for opening, without being enclosed, is put in the CDO section of an AIP of the destination archive. The description language of the manual is given in the RI section. The PDI will record “the manual for opening the item deposited by OAIS S.” If programs for opening and disclosure have been prepared in advance, they are put in the CDO section. Environment information for executing the programs is put in the RI section. The archive systems guarantee that any AIP transferred to a destination archive can be rebuilt as it was but they cannot guarantee that the programs enclosed in an AIP are executable at any time.

Return work - The enclosing document returned from the destination archive is opened by the source system and the recovered AIP is registered to the source system. Opening work is performed by referring to the manual for opening.

Disclosure work – In the case that the destination archive is requested to recover an AIP transferred from a source archive, the archive discloses the enclosing document, opens it as required, and registers it to the destination system.

5. Model for Inter-Archival Deposit of an Archived Collection

"Collection" is a unit to manage a set of resources collected based on a certain policy. A collection of resources is a natural unit for preservation. In the OAIS reference model, collections are expressed as follows. There are two types of AIP: one is an Archival Information Collection (AIC) and the other is an Archival Information Unit (AIU), which can be a member of the former. The CI of an AIU holds a single item of archival information. The CI of an AIC will hold information that enables identifying individual AIUs as members of the collection.

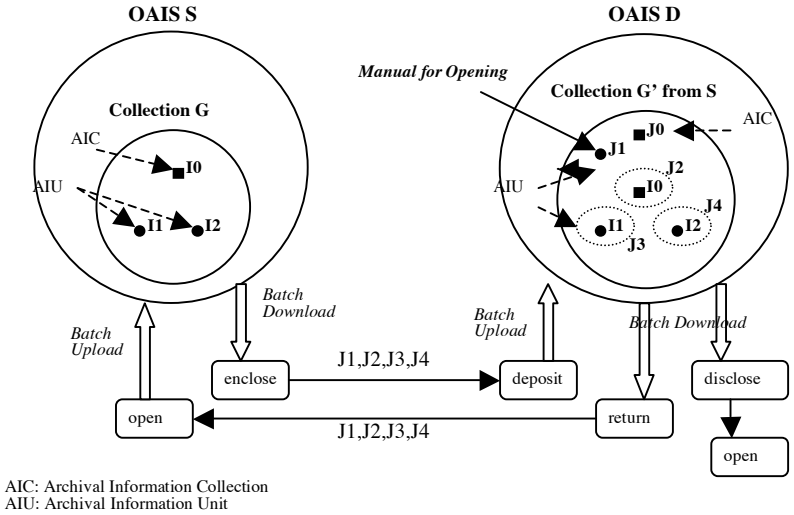

Fig. 4. Depositing a Collection between OAISs

(1) Enclosing and depositing

In Fig. 4, the source side which has a collection G prepares a series of enclosing documents J2, J3, and J4 for the AIC and AIUs. The enclosing documents J2, J3, and J4, together with a manual for opening J1, are transferred to the destination side. Then, the destination side creates AIPs (as AIUs of a collection G') from J2, J3, J4 and J1. In this process, the download/upload functions may be realized as a batch process.

(1.1) Source side: Download and Enclose

Because both the AIC and the AIU are AIPs, the process of preparing enclosing documents is the same as described in Section 4. These are downloaded as a batch from the source archive and enclosing work is performed item by item. A manual for opening applicable to all enclosing documents is prepared as described in Section 4.

(1.2) Destination side: Deposit and Upload

On the destination side, the process of creating destination AIPs from the enclosing documents transferred from the source side is the same as described in Section 4. Since a collection has a number of items, this process is realized as a batch process as well. When the destination side creates the AIPs, it assigns a unique identifier to each of their PDIs. The destination archive forms a collection G' by considering the batch-uploaded AIPs to be AIUs. The manual for opening J1 is added to this collection, also as an AIU. The AIC for this collection, i.e., J0 is prepared by the destination archive. The manual for opening and the AIC may be individually registered by operator's commands.

(2) Return and Opening

The destination archive identifies the deposited collections and downloads them as a batch. Enclosing documents J2, J3, J4 are extracted from all CDO sections in the AIPs of the destination archive and are returned to the source system, together with the manual for opening J1. The AIPs (AIC and AIUs) of the source side are extracted from a series of enclosing documents, referring to the manual for opening and any opening programs that may be attached. The AIPs are then uploaded as a batch to the source archive to recover the collections.

(3) Disclosing and Opening

The destination archive identifies the deposited collections and downloads them in a batch process. Enclosing documents are extracted from all CDO sections in the AIPs of the destination archive. The enclosing documents (XML documents) are arrayed into a single XML document as follows.

```
<?xml version="1.0" encoding="xxx" ?>
<Collection>
<AIP><CI><CDO>yy0</CDO><RI>xxx</RI></CI><PDI>xxx</PDI></AIP>
<AIP><CI><CDO>yy1</CDO><RI>xxx</RI></CI><PDI>xxx</PDI></AIP>
<AIP><CI><CDO>yy2</CDO><RI>xxx</RI></CI><PDI>xxx</PDI></AIP>
</Collection>                                                    (2)
```

CDO	RI	PDI
C:/ppp/qqq/1/rrr0.xml	xxx	xxx
C:/ppp/qqq/2/rrr1.xml	xxx	xxx
C:/ppp/qqq/3/rrr2.xml	xxx	xxx

CDO: **Location of the enclosing document** instead of data objects

Fig. 5. Table of Source AIPs obtained by Disclosing:

For the management of archiving a collection by the collaborating archives, we need a management tool. In the implementation discussed in the next section, we use Microsoft Excel to view the list of objects in Expression (2) as shown in Fig.5. In Fig.5, the elements in column CDO are pointers to indicate the locations of enclosing documents in the downloaded files.

6. A Pilot System of the Collaborative Archive Model using DSpace

To verify the proposed model, we have configured a pilot system (see Fig. 6). DSpace [3] is the archiving system, since it is widely used and fulfils the basic feature of the OAIS reference model. Windows XP is the environment for enclosing, depositing, returning, opening, and disclosing.

This DSpace-to-DSpace system is not a simple back-up mirroring. In the destination system, a resource deposited by the source system cannot be accessed directly by an end-user because it is enclosed in a bag. Furthermore, the function implemented in the destination DSpace is valid for any architecture of source system because the destination system merely receives a bag (i.e., an XML document) from it. Similarly, the function implemented in the source DSpace is valid for any architecture of destination system because the bag that was deposited is merely returned.

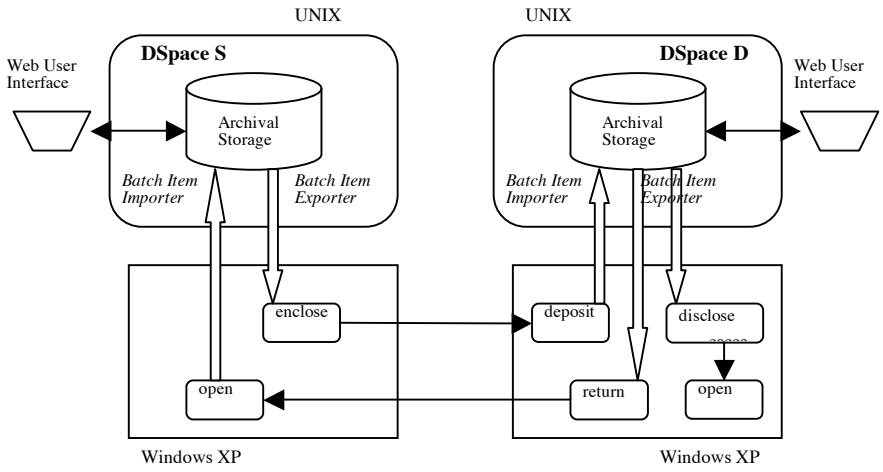

Fig. 6. Verification of the Model using DSpace

During the demonstration, enclosing, depositing, returning, opening, and disclosing were performed in units of collection. DSpace has a Web interface for item-by-item processing and a command for batch processes. Batch Item Exporter and Batch Item Importer are commands for batch downloading and batch uploading. Both specify a collection and then process its elements as a batch. These batch commands use the DSpace simple archive format [4], which is a directory full of items with a subdirectory per item. Each item directory contains files including AIP components. Here, file extensions correspond to the RI.

Each enclosing, depositing, returning, opening, and disclosing function was prepared based on the above process and was implemented under Windows XP, as depicted in Fig. 6. Microsoft Excel and Microsoft Access were used for disclosure.

7. Discussion and Conclusion

This study explored a simple approach to making archives more reliable. Such simple technology is indispensable for archives maintained by memory organizations of small communities, such as regional libraries and museums. On the one hand, each community needs its own policies to archive and preserve resources, which are valuable not only for the local community but also for the global community. On the other hand, these small communities are subject to change or may disappear over time.

The only solution is to archive the resources in another archive. However, such multiple archiving involves the challenges of intellectual property issues and interoperability between archives. Intellectual property issues are beyond the scope of this paper. Interoperability between archives is crucial in the heterogeneous Internet environment. It is unrealistic to assume that all archives use a single archiving software or that all archives use the same metadata schema. Thus, interoperability between archives is difficult to achieve.

A simple, open framework is required to solve the interoperability issue. The proposed model is simple: enclose an AIP in a bag and store the bag packaged in another AIP at a different archive. The model is open: the enclosing document format (i.e., AIP transfer format) is defined based on XML. In this model, the intellectual property issue is partially solved by restricting an end-user's access to the deposited resources on the destination side.

In the destination archive, migration of the deposited resource is not difficult because it is merely a text object such as an XML document. However, when migration occurs in the source archive, the resource needs to be re-enclosed and re-deposited in the destination archive.

Many archives are not as reliable as those maintained by large memory organizations, but their resources should be preserved as well as those of the large ones. The simplicity of the proposed model is a key to solving this issue.

Acknowledgements

This work is supported by JSPS Grants-in-Aid for Scientific Research (JSPS: Japan Society for the Promotion of Science).

References

1. LOCKSS Web Site, http://lockss.stanford.edu/index.html
2. Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002. This Recommendation has been adopted as ISO 14721:2003, http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf
3. Robert Tansley, et al; DSpace as an Open Archival Information System: Current Status and Future Directions, ECDL2003, Lecture Notes in Computer Science 2769, pp446-460, 2003
4. Robert Tansley, et al; DSpace System Documentation, 2004, http://dspace.org/technology/system-docs/