2004

Resource Harvesting Within the OAI-PMH Framework

Herbert Van de Sompel
Los Alamos National Laboratory

Michael L. Nelson
Old Dominion University

Carl Lagoze
Cornell University

Simeon Warner
Cornell University

Original Publication Citation

Van De Sompel, H., Nelson, M.L., Lagoze, C., & Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-Lib Magazine, 10*(12). doi: 10.1045/december2004-vandesompel

ARTICLES

D-Lib Magazine
December 2004

Volume 10 Number 12

ISSN 1082-9873

Resource Harvesting within the OAI-PMH Framework

Herbert Van de Sompel
Los Alamos National Laboratory, Research Library
<herbertv@lanl.gov>

Michael L. Nelson
Old Dominion University, Computer Science Department
<mln@cs.odu.edu>

Carl Lagoze
Cornell University, Computing and Information Science <lagoze@cs.cornell.edu>

Simeon Warner
Cornell University, Computing and Information Science
<simeon@cs.cornell.edu>

Abstract

Motivated by preservation and resource discovery, we examine how digital resources, and not just metadata about resources, can be harvested using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). We review and critique existing techniques for identifying and gathering digital resources using metadata harvested through the OAI-PMH. We introduce an alternative solution that builds on the introduction of complex object formats that provide a more accurate way to describe digital resources. We argue that the use of complex object formats as OAI-PMH metadata formats results in a reliable and attractive approach for incremental harvesting of resources using the OAI-PMH.

Introduction

The Open Archives Protocol for Metadata Harvesting (OAI-PMH) [Lagoze et al. 2002] has been widely adopted as an approach to allow harvesting of metadata. This metadata pertains to resources that are themselves outside of the scope of the OAI-PMH data model. The exposed metadata is typically of a descriptive nature, and is expressed by means of metadata formats of varying complexity, such as Dublin Core [note 1], or MARCXML [note 2].

Recently, use cases have emerged that reveal a more liberal interpretation of what constitutes metadata in the OAI-PMH [Van de Sompel, Young and Hickey, 2003]. For example, in a joint project of the Los Alamos National Laboratory (LANL) and Old Dominion University, the OAI-PMH is used as a means to make usage information pertaining to digital resources harvestable. In that project, each exposed metadata record summarizes the access history for a specific digital resource.

In this paper we retain the view of metadata as being descriptive. However, we expand the scope of descriptive metadata to be more than just DC, MARC and similar bibliographic formats. This allows the introduction of metadata formats that are more complex, expressive, and accurate in their description of digital resources. Such formats exist, and are generally referred to as *complex object formats*; examples include MPEG-21 DIDL [note 3], METS [note 4], and SCORM [note 5]. The combination of these complex object formats with the OAI-PMH results in a framework that allows for reliable harvesting of digital resources.

Problem Statement

There is a growing need to make resources, not only descriptive metadata, harvestable in an interoperable manner. There are two major use cases that motivate this need:

- Preservation: The need to periodically transfer digital content from a data repository to one or more trusted digital repositories charged with storing and preserving safety copies of the content. The trusted digital repositories need a mechanism to automatically synchronize with the originating data repository.
- Discovery: The need to use content itself in the creation of services. Examples include search engines that make full-text from multiple data repositories searchable, and citation indexing systems that extract references from the full-text content. Another scenario is the provision of thumbnail versions of high-quality images from cultural heritage collections to external services that build browsing interfaces that include the thumbnails [Osborne 2004].

Both the preservation and discovery use cases have been discussed in the context of Digital Library and Institutional Repository projects in The Netherlands [note 6], the UK [note 7] and Germany [note 8]. The preservation use case is also emerging in the Archive Export/Ingest effort of the National Digital Information Infrastructure and Preservation Program [note 9]. The discovery use case has also emerged in the realm of web search engines, where both the sophistication of search technology and content coverage are competitive factors. This has led to growing interest by search engine providers in "deep web" content stored in digital libraries and institutional repositories, as exemplified by collaborations between OAIster [note 10], OCLC [note 11], arXiv [note 12], NSDL [note 13] and major web search engines. It is thus intriguing to consider if the widespread adoption of the OAI-PMH could be used as leverage to address the content gathering requirement.

A number of methods already exist for indirectly gathering digital resources through the metadata exposed by OAI-PMH repositories. In most cases, the Dublin Core metadata exposed by these repositories describes digital resources available at a network location. These digital resources, and the network location from which they are available, are typically under control of the data provider that operates the OAI-PMH repository. A common approach to content gathering in these cases is as follows:

1. An OAI-PMH harvester harvests Dublin Core records from the OAI-PMH repository.
2. The harvester analyzes each Dublin Core record, extracting dc.identifier information in order to determine the network location of the described resource.
3. A separate process, out-of-band from the OAI-PMH, collects the described resource from its network location.

The two most significant issues that arise with this approach are:

- **Locating the resource based on information provided in dc.identifier.** This problem actually comes in two guises, and has been reported to be a source of considerable frustration [Lossau, 2004; Summann & Lossau, 2004]:

- dc.identifier is commonly, and legitimately, used to convey a variety of resource identifiers [note 14]. The element may simultaneously include a URL, a DOI, a bibliographic citation, etc. Since dc.identifier does not have the expressiveness to unambiguously convey which of the provided identifiers, if any, is a locator for the resource, a variety of heuristics and several dereferencing attempts have been used to try to locate and gather the resource.
- Because Dublin Core records are typically created for provision of user services, in many cases a network location provided in dc.identifier is that of a page that provides a link to the resource—a so-called splash-page—not that of the resource itself. In these cases, issues that arise include determining whether the dereferenced object is actually the resource or a proxy for it, and—if it is a proxy—trying to find the embedded link that leads to the actual resource.

- **OAI-PMH datestamp to trigger resource harvesting.** When gathering content, a harvester must unambiguously know when digital resources have been added or modified, since synchronization between a data repository and the services that use its content is essential. In the typical OAI-PMH scenario described above, the available OAI-PMH datestamp is, by definition, the date and time of creation or modification of the Dublin Core metadata record. Modifying a resource does not necessarily yield a modification of the associated Dublin Core record. Thus, the OAI-PMH datestamp is not a reliable basis for date-based harvesting of resources. Table 1 shows the problematic issues that arise when the OAI-PMH datestamp of the Dublin Core record is used as an indication of the creation or modification time of the described resource.

	no metadata update	**metadata update**
no resource update	OK	unnecessary resource download
resource update	missed resource update	OK

Table 1: Issues arising when the OAI-PMH datestamp, which changes only when the metadata is updated, is used as a basis for content gathering

Existing Approaches

The need for incremental harvesting of resources has led to a number of techniques aimed at overcoming the described issues. All the techniques proposed have problems that make them undesirable as a general solution to the problem.

One example is the mirroring agreement between the NACA Technical Report Server [Nelson, 1999] and the MAGiC project in the UK [Sidwell, Needham & Harrington, 2000]. Although each collection consisted of historical aeronautical reports, the reports were being scanned and added to the collection at bursty, uneven rates, from which arose the need to synchronize the sites. Because both sites had OAI-PMH repositories it was decided to use the OAI-PMH interface to learn of new additions to each organization's collection. The OAI-PMH datestamp of metadata records was used

to trigger harvesting of resources (i.e., PDFs). The nature of the resources made it very unlikely that the resources would be updated without the metadata being updated, thereby greatly decreasing the risk of missed updates as indicated in Table 1. However, this mirroring solution required significant point-to-point human communication and ad-hoc solutions due to a number of problems:

- Although the MAGiC metadata records list the actual URL of the digital resource (i.e. the PDF) in the dc.identifier element, the NACA metadata records list the URL of a bibliographic splash-page in dc.identifier. Mirroring of the NACA reports therefore required ad-hoc construction of the resource URL from the URL in dc.identifier.

- The NACA reports contained much more than just the PDF; they also contained GIFs, TIFFs, and text files resulting from Optical Character Recognition of TIFF files. While the latter two files types are Web accessible, links to them are not directly exposed from the bibliographic splash-page specified in the NACA dc.identifier.

These issues stand in the way of making the NACA/MAGiC mirroring approach generalizable from an archiving perspective.

Implementers have proposed a number of methods for expressing the network location of the digital resource within the DC metadata record. One conventional heuristic is to use the first resolvable dc.identifier element of a Dublin Core record as the URL of the resource [Young, 2004]. Another solution (Example 1) is to use the dc.format element as a means for conveying the network location of the resource. As can be seen, the URL is preceded by an indication of the MIME subtype [Freed & Borenstein, 1996] of the resource. Another approach (Example 2) conveys the URL of the resource in the dc.relation element. And in yet another approach (Example 3), the URL of the resource is also in the dc.relation element, but there is a second dc.relation element that specifies an alternate URL for the bibliographic splash-page that is also specified in dc.identifier.

All three approaches exist in eprints.org [note 15] installations. Examples 1-3 illustrate these approaches for the description of the same resource. Example 1 is directly taken from an OAI-PMH repository, while Examples 2 and 3 are mock-ups that illustrate techniques used by other repositories. The lack of a general mechanism, even in the context of a single OAI-PMH software implementation, is a significant burden for applications that require the gathering of content, not only metadata.

```
<oai_dc:dc   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
 http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
 xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/">
    <dc:title>A Simple Parallel-Plate Resonator Technique for Microwave.
         Characterization of Thin Resistive Films</dc:title>
    <dc:creator>Vorobiev, A.</dc:creator>
    <dc:subject>ING-INF/01 Elettronica</dc:subject>
    <dc:description>A parallel-plate resonator method is proposed for
          non-destructive characterisation of resistive films used in microwave
          integrated circuits. A slot made in one ...
       </dc:description>
    <dc:publisher>Microwave engineering Europe</dc:publisher>
    <dc:date>2002</dc:date>
    <dc:type>Documento relativo ad una Conferenza o altro Evento</dc:type>
    <dc:type>PeerReviewed</dc:type>
    <dc:identifier>http://amsacta.cib.unibo.it/archive/00000014/</dc:identifier>
    <dc:format>pdf
        http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf
       </dc:format>
```

```
</oai_dc:dc>
```

Example 1: The URL of the resource is specified in dc.format, prefixed with "pdf". The URL of the bibliographic splash-page is provided in dc.identifier.

```
...
    <dc:identifier>http://amsacta.cib.unibo.it/archive/00000014/
      </dc:identifier>
    <dc:relation> http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf
      </dc:relation>
...
```

Example 2: The URL of the resource is specified in dc.relation. The URL of the bibliographic splash-page is provided in dc.identifier.

```
...
    <dc:identifier>http://amsacta.cib.unibo.it/archive/00000014/
      </dc:identifier>
    <dc:relation>http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf
      </dc:relation>
    <dc:relation> http://resolver.unibo.it/00000014/
      </dc:relation>
...
```

Example 3: The URL to both the resource and the bibliographic splash-page are specified in dc.relation. A different URL for the bibliographic splash-page is provided in dc.identifier.

Another approach that has been proposed is to convey the URL of a bibliographic splash-page in the dc.identifier, and, to specify the URL(s) of the resource in special-purpose XHTML <link> elements [Tourte & Powell, 2004]. This approach assumes the existence of a splash-page and requires a harvester to be able to determine that it is harvesting from an OAI-PMH repository that follows the proposed convention. Furthermore, the possibility of missed updates to the resources remains unaddressed because the OAI-PMH datestamp of the metadata record remains as the supposed indicator of content modification.

This mixture of approaches suggests that unqualified Dublin Core does not possess sufficiently rigorous semantics to unambiguously express the information essential for resource harvesting. In response, implementers have explored the use of more expressive, qualified Dublin Core format to address the issues. Indeed, qualified Dublin Core has the expressiveness to disambiguate between various types of identifiers and locators. However, it is unclear how use of a qualified Dublin Core format would resolve the issues described earlier regarding the OAI-PMH datestamp. These issues must be resolved as part of a reliable content gathering solution. Since more expressive complex object formats exist, and are being accepted and deployed, it seems constructive to pursue that path instead of trying to force-fit the solution into a metadata format ill-suited for such purpose. This is the solution we propose in the remainder of this paper.

At least one project attempts to address the issue in a more fundamental manner. The OA-X effort [note 16] extends the OAI-PMH with an extra verb aimed solely at gathering content. While protocol extensions can be successfully deployed in limited contexts, they frequently fail to be generalizable. When new use cases arise, such as resource harvesting, it is prudent to first and thoroughly explore approaches that do not make obsolete or fracture existing OAI-PMH installations.

A Solution within the OAI-PMH Framework: Complex Object Formats as OAI-PMH Metadata Formats

The above examples show that it is possible to harvest resources using extensions or ad-hoc conventions outside of the OAI-PMH. Each of these techniques has problems that interfere with their generality and specificity. While these problems may be merely frustrating for certain applications,

they are unacceptable for others. This is certainly true for preservation, where the goal is to create perfectly synchronized archives of a given data repository. The problems also interfere with resource harvesting by discovery services that rely on an accurate reflection of repository content in their indexes.

In order to accommodate these more demanding use cases within the boundaries of the OAI-PMH, we introduce the use of metadata formats that are more complex, expressive, and accurate in their description of digital resources. These formats have specifically been defined to represent digital objects [Kahn & Wilensky, 1995].

Complex Object Formats

Expressive formats that permit representation of digital objects have emerged from several communities, and are commonly referred to as *complex object formats*. A historical survey of complex objects is provided by [Nelson et al., 2001] and recent examples include MPEG-21 DIDL [note 3], METS [note 4], and SCORM [note 5]. Complex object formats typically share the following core characteristics:

- Representation of a digital object by means of a wrapper XML document.
- The ability to represent both simple digital objects (consisting of a single datastream), and compound digital objects (consisting of multiple datastreams).
- The ability to unambiguously convey identifiers of the digital object and its constituent datastreams.
- The ability to include a datastream in two, not mutually exclusive, ways:
    - By-Value: embedding a base64-encoding [Freed & Borenstein, 1996] of the datastream inside the wrapper XML document.
    - By-Reference: unambiguously embedding the network location of the datastream inside the wrapper XML document. This approach is considered fully equivalent with the By-Value approach.
- The ability to include a variety of secondary information pertaining to a datastream. This includes descriptive metadata, rights information, technical metadata, etc. This secondary information can also be provided By-Value or By-Reference.

The OAI-PMH Data Model and Complex Object Formats

Figure 1 depicts the OAI-PMH data model for cases where the resource, as defined in the OAI-PMH data model, is digital content. It introduces this expanded view of descriptive metadata. The figure should be interpreted as follows (OAI-PMH concepts are written in *italic*):

- At the very top is a digital *resource* (a PDF file, for example) about which an OAI-PMH repository exposes *metadata*. Note that the digital *resource* can also be compound, i.e. consist of multiple datastreams. As was previously mentioned, *resources* themselves are outside of the scope of the OAI-PMH.
- Listed below the *resource* is the *item*. The *item* is the highest-level entity within the scope of the OAI-PMH. In essence, the *item* is the entry point to all available *metadata* pertaining to a *resource*. In the protocol, the *item* is uniquely identified by an OAI-PMH *identifier*.

- Below the *item*, several records are shown. *Records* contain *metadata* (and secondary information about that *metadata*). A specific record in the OAI-PMH is unambiguously identified by means of the combination of the OAI-PMH *identifier* (of the *item*), the *metadataPrefix* that specifies the *metadata format* used for the dissemination of the *metadata*, and the OAI-PMH *datestamp* of the *metadata*. The *datestamp* is the date and time of creation or modification of *metadata*. Note that the *datestamp* is a property of the *metadata* record, not of the *item* as used to be the case in previous protocol versions [Lagoze et al, 2002]. This reflects the fact that *metadata* of various *metadata formats* may be made available and may be modified independently, thus having different *datestamps*.

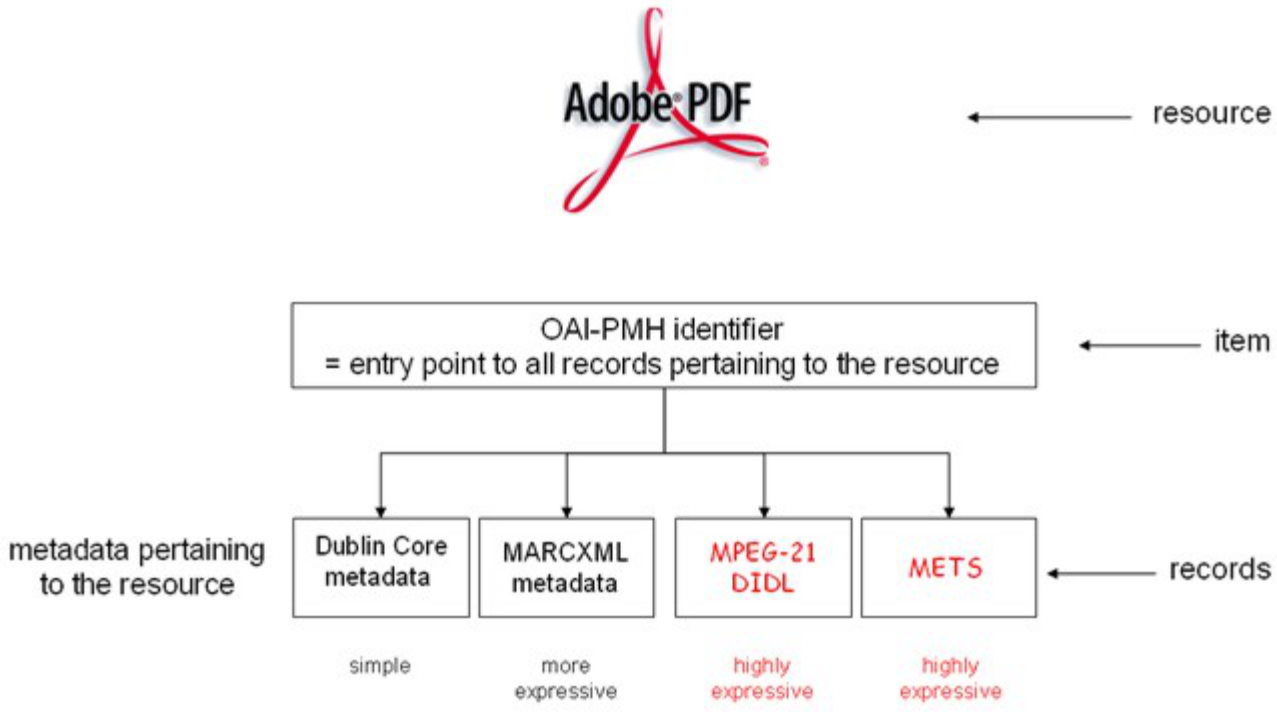

Figure 1: The OAI-PMH data model

Figure 1 introduces complex object formats in the OAI-PMH data model. The figure depicts descriptive metadata with an increasing degree of complexity and accuracy. Dublin Core metadata is a description of the resource according to the minimalist, resource discovery-oriented, Dublin Core metadata format. MARCXML metadata describes the resource according to the more complex, expressive and cataloging-oriented, MARC metadata format. And, an MPEG-21 DIDL XML document or a METS XML document describes the resource according to an even more complex format that focuses on the accurate representation of digital objects. These formats allow expression of a variety of secondary information pertaining to the resource, including descriptive, rights, technical, structural, and provenance metadata. They also allow unambiguously conveying identifiers, and inclusion of the resource itself By-Reference or even By-Value. These formats can be used to represent digital objects of all kinds; they can represent digital objects irrespective of the type and number of contained datastreams. It is worthwhile noting that the ability to describe resources of various classes is a core characteristic of metadata formats; it is a characteristic that distinguishes metadata formats from file formats. Based on these considerations, it is legitimate to consider complex object formats as metadata formats. This perspective is also supported by the name METS itself: *Metadata* Encoding and Transmission Standard.

Example 4 shows a representation of a resource by means of the MPEG-21 Digital Item Description

Language (DIDL). The resource is the same as that described with Dublin Core metadata in Examples 1-3. Key features of the DIDL representation of the resource are:

- The resource is mapped to the didl:Item entity of the DIDL data model.
- The didl:Item has two didl:Descriptors that convey secondary information about the resource:
    - The first didl:Descriptor builds on the MPEG-21 Digital Item Identification Standard [note 17] to unambiguously convey the identifier (http://amsacta.cib.unibo.it/archive/00000014/) of the resource.
    - The second didl:Descriptor conveys Dublin Core metadata pertaining to the resource that is almost identical to that conveyed in Example 3. The only difference is that the ambiguous specification of the network location of the resource using the dc.relation field has been removed.
- The network location of the resource (http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf) is provided unambiguously as the value of the @ref attribute of the didl:Resource element. In addition to that, the @mimeType attribute of the same element specifies the mime type of the referenced resource in a standard-based manner.

```
<didl:DIDL  xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
             xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
             xsi:schemaLocation="urn:mpeg:mpeg21:2002:02-DIDL-NS
             http://purl.lanl.gov/STB-RL/schemas/2004-11/DIDL.xsd">
 <didl:Item>
     <didl:Descriptor>
         <didl:Statement mimeType="text/xml; charset=UTF-8">
             <dii:Identifier
                     xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
                     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                     xsi:schemaLocation="urn:mpeg:mpeg21:2002:01-DII-NS
                     http://purl.lanl.gov/STB-RL/schemas/2003-09/DII.xsd">
                http://amsacta.cib.unibo.it/archive/00000014/
             </dii:Identifier>
         </didl:Statement>
     </didl:Descriptor>
     <didl:Descriptor>
         <didl:Statement mimeType="text/xml; charset=UTF-8">
            <oai_dc:dc xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
                xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
                xmlns:dc="http://purl.org/dc/elements/1.1/">
             <dc:title>A Simple Parallel-Plate Resonator Technique for Microwave.
                Characterization of Thin Resistive Films
             </dc:title>
             <dc:creator>Vorobiev, A.</dc:creator>
             <dc:subject>ING-INF/01 Elettronica</dc:subject>
             <dc:description>A parallel-plate resonator method is proposed for
                non-destructive characterisation of resistive films used in microwave
                integrated circuits. A slot made in one ...
             </dc:description>
             <dc:publisher>Microwave engineering Europe</dc:publisher>
             <dc:date>2002</dc:date>
             <dc:type>Documento relativo ad una Conferenza o altro Evento</dc:type>
             <dc:type>PeerReviewed</dc:type>
             <dc:identifier>
                http://amsacta.cib.unibo.it/archive/00000014/
             </dc:identifier>
```

```
            <dc:format>application/pdf</dc:format>
          </oai_dc:dc>
        </didl:Statement>
    </didl:Descriptor>
    <didl:Component>
        <didl:Resource mimeType="application/pdf" ref="http://amsacta.cib.unibo.it/archive/
          00000014/01/GaAs_1_Vorobiev.pdf"/>
    </didl:Component>
</didl:Item>
</didl:DIDL>
```

Example 4: The resource of Example 1-3 described according to the MPEG-21 Digital Item Description Language.

Example 4 shows just one of the many representations that could be devised using MPEG-21 DIDL. MPEG-21 DIDL and other complex object formats are quite versatile and can be used to construct various representations of the same resource. However, the common set of core features noted above enables all the complex object formats mentioned to support:

- Descriptive metadata formats other than Dublin Core.
- Resources consisting of multiple datastreams. These are still represented in a single description but each datastream can have an unambiguous network location associated with it. Multiple datastreams may also have their own identifiers, share identifiers, or both. The same applies to other secondary information about such datastreams, including descriptive metadata.
- By-Value inclusion of resources (or datastreams) using base64-encoding. While this technique comes with technical challenges when large datastreams are concerned, it is clearly attractive for reasonably sized datastreams. For example, providing thumbnails of images By-Value is clearly feasible.

Use of Complex Object Formats with the OAI-PMH

The combination of complex object formats and the OAI-PMH is an attractive option to address the content gathering issues discussed so far. Table 2 illustrates how the complex object metadata format fits into the OAI-PMH data model. Attractive features of this approach include:

- Complex object formats represent a resource by means of a wrapper XML document, and therefore, a representation can natively be conveyed as metadata (inside the <metadata> element) of OAI-PMH responses.
- Complex object formats provide the expressiveness to unambiguously specify the network location of the resource (and/or its constituent datastreams). This solves the problem of locating the resource that is to be harvested.
- Use of a complex object format as a separate metadata format within in the OAI-PMH framework yields an unambiguous trigger mechanism for harvesting resources. Indeed, per definition, the OAI-PMH datestamp is the date of creation or modification of metadata. When using a complex object format, that metadata is a representation of the resource that covers all its constituents including its multiple datastreams, its descriptive metadata, etc. As a result, as soon as a change occurs in one of these constituents, the associated OAI-PMH datestamp must change. It should be noted that such changes do not necessarily result in a change of the wrapper XML document. Indeed, if changes occur to a bitstream that is provided By-Reference, its network location may remain unchanged, and hence, the wrapper XML document may remain unchanged.

However, because the By-Value and By-Reference provision of bitstreams are considered equivalent in complex object formats, the OAI-PMH datestamp must change in the By-Reference approach, as it would in a By-Value approach. As a result, the addition of new resources or the modification of existing resources can be detected using the OAI-PMH datestamp of the complex object representation of the resource.

- Complex object formats provide a uniform resource harvesting solution both when the resource is simple (a single datastream) and when it is compound (consisting of multiple datastreams).
- Complex object formats provide the ability to disambiguate between identifiers and locators of resources (or datastreams), and even to disambiguate between identifiers and locators of resources and those of metadata
- When complex object formats are used within OAI-PMH, properties such as set-membership and the use of "about" containers (to convey secondary information pertaining to metadata) apply with consistent semantics to metadata records that are complex object representations just as they do to other metadata records.

OAI-PMH Entity	Value	Description
Resource	URL	PDF, PS, XML. HTML or other file
Item		
identifier	oai-identifier	Identifier that follows oai-identifier scheme [Lagoze et al., 2002]
set membership	LCSH	Library of Congress Subject Heading
Record 1		
metadataPrefix	oai_dc	bibliographic metadata in Dublin Core format
datestamp	2004-07-22	modification date of the Dublin Core record
Record 2		
metadataPrefix	marc21	bibliographic metadata in MARC format
datestamp	2004-07-31	modification date of the MARC record
Record 3		
metadataPrefix	didl	Representation of the resource using MPEG-21 DIDL
datestamp	2004-08-02	modification date of the last modified constituent of the resource

Table 2: An OAI-PMH data model perspective of an OAI-PMH repository supporting 3 metadata formats: Dublin Core, MARCXML, and MPEG-21 DIDL.

A typical scenario for resource harvesting using complex object formats within the OAI-PMH would be:

1. The OAI-PMH harvester checks for support of a locally understood complex object format

using the ListMetadataFormats verb.

2. When support is detected, the harvester harvests the complex object metadata from the repository. Semantics of the OAI-PMH datestamp for records in this format guarantee that new and modified resources are detected.
3. A parser at the end of the harvesting application analyzes each harvested complex object record:
    - The parser extracts the bitstreams that were delivered By-Value.
    - The parser extracts the unambiguous references to the network location of bitstreams delivered By-Reference.
4. A separate process, out-of-band from the OAI-PMH, collects the bitstreams delivered By-Reference from the extracted network locations.

The OAI-PMH and Complex Object Formats: Existing Implementations

Various projects are already exploring the use of complex object formats in combination with the OAI-PMH. They are described in this section.

LANL repository

At the LANL Research Library, digital objects are represented using MPEG-21 DIDL, and function as Archival Information Packages (AIPs) [note 18] in the Digital Library repository. The OAI-PMH is used as a repository access protocol. The ListRecords verb is used by downstream applications to incrementally harvest content on a datestamp/set basis. The GetRecord verb is used to request the dissemination of a single AIP. This approach has been used in production since June 2004, and research leading to the approach is documented in [Bekaert, Hochstenbach & Van de Sompel, 2003; Bekaert et al., 2004; Jerez, et al. 2004]. At the time of writing, the LANL repository contains 15,000,000 AIPs, a figure that is expected to triple in the next 12 months. Off-the-shelf OAI-PMH tools such as OCLC's OAICat [note 19], OAIHarvester [note 20] and OAI Viewer [note 21] are used throughout the repository infrastructure. Example 5 illustrates the use of the MPEG-21 DIDL and the OAI-PMH in the LANL Repository.

OAI-PMH GetRecord response from the LANL Repository that contains a DIDL representation of a resource (A BIOSIS record): LANL_GetRecord.xml
Movie 1: An interaction with a constituent OAI-PMH repository of the LANL Repository: • Camtasia Pack & Show executable for WinTel computer; no audio; size = 4.9 Mb: LANL_OAIPMH.exe • QuickTime movie; no audio; size = 15 Mb: LANL_OAIPMH.mov

Example 5: Illustrations of the use of complex object formats in the LANL Repository.

Mirroring the collection of the American Physical Society at LANL

Another project ongoing at the Research Library of the Los Alamos National Laboratory aims at accurate and timely mirroring of the collection of the American Physical Society (APS) by means of the OAI-PMH. In this application, mirroring refers to duplication of APS content at LANL, not to mirroring the APS application, nor to duplicating the lower-level storage and repository approach used by the APS. A digital object created by the APS typically consists of multiple datastreams,

including expressive descriptive metadata, a research paper in various formats, and auxiliary content such as datasets, video recordings, etc. In the project, each such object is exposed as a complex object through the already existing, limited-access, APS OAI-PMH interface. Again, MPEG-21 DIDL is used as the complex object format, and the APS has integrated a module into the OAI-PMH interface that facilitates representing their digital objects as DIDL XML documents. Through periodic OAI-PMH harvesting, LANL collects updated and added content from the APS, thereby relying on the semantics of the OAI-PMH datestamp for exposed complex objects to ensure accurate duplication of content. As is the case with typical metadata harvesting, it is not expected that the whole 690 GB APS repository will be harvested using the OAI-PMH. Rather, it is anticipated that the APS archive, until a specified date, will be delivered out-of-band, using physical media. Starting from that date, updates to the APS repository will be harvested using the OAI-PMH. At the time of writing, results from ongoing experiments suggest that the tested approach can eventually be successfully brought into production. Example 6 illustrates the use of the MPEG-21 DIDL and the OAI-PMH in the APS/LANL mirroring project.

OAI-PMH GetRecord response from the APS Repository containing a DIDL representation of a resource (an APS publication): APS_GetRecord.xml
Movie 2: An interaction with the OAI-PMH repository of the APS: • Camtasia Pack & Show executable for WinTel computer; no audio; size = 4.3 Mb: APS_OAIPMH.exe • QuickTime movie; no audio; size = 16 Mb: APS_OAIPMH.mov

Example 6: Illustrations of the use of complex object formats in the APS/ LANL mirroring project.

DSpace and Fedora plug-ins

In an attempt to more publicly demonstrate the possibility of harvesting content within the boundaries of the OAI-PMH, the LANL Research Library has created an experimental plug-in for DSpace v.1.2 systems [note 22] that allows DSpace [note 23] items to be exposed as MPEG-21 DIDL XML documents via the existing DSpace OAI-PMH interface. Through the installation of this plug-in, system administrators can facilitate experimentation with mirroring the content of DSpace repositories or with creating services based on DSpace content, not only metadata. The plug-in allows administrators to specify the maximum size of the bitstreams that will be delivered By-Value, allowing them to control the harvesting load in their environment. When setting the maximum size to zero, all bitstreams are delivered By-Reference, avoiding potential scalability problems at the harvesting side. A project aimed at creating a similar, experimental plug-in for Fedora [note 24] repositories has also been launched. Example 7 illustrates the use of the MPEG-21 DIDL and the OAI-PMH in a DSpace repository for which the DSpace DIDL plug-in was installed.

OAI-PMH GetRecord response from a DSpace repository containing a DIDL representation of a resource (a DSpace item): DSpace_GetRecord.xml
Movie 3: An interaction with the OAI-PMH interface of a DSpace repository for which the DIDL plug-in was installed: • Camtasia Pack & Show executable for WinTel computer; no audio; size = 3.3 Mb: DSpace_OAIPMH.exe • QuickTime movie; no audio; size = 14.6 Mb: DSpace_OAIPMH.mov

Example 7: Illustrations of the use of complex object formats with the OAI-PMH interface to a DSpace repository.

mod_oai

mod_oai is a Mellon-funded joint project between Old Dominion University and the LANL Research Library that aims to bring OAI-PMH semantics to the web crawling community [note 25]. An Apache module, mod_oai, is being developed that automatically responds to OAI-PMH requests on behalf of a web server. If Apache and mod_oai are installed at:

http://www.foo.edu/

then the baseURL for the OAI-PMH repository is:

http://www.foo.edu/mod_oai

While respecting the http access controls specified in httpd.conf, mod_oai provides 3 metadata formats in the OAI-PMH responses. Dublin Core is provided, but only technical metadata such as file size and MIME type is included. A new metadata format is introduced, http_header, which contains all the http response headers that *would have been returned* if the resource had been obtained by a regular Web crawler. The third metadata format is oai_didl, which represents the Web resource according to the MPEG-21 DIDL format. This representation includes the metadata in the http_header format, as well as the Web resource itself, provided using the By-Reference or By-Value approach, or both.

There are two general classes of mod_oai use. The first is to issue only ListIdentifiers as a way of identifying new URLs to be added to a regular web crawler. The second is to use ListRecords to retrieve the resources in the oai_didl format. Example 8 illustrates the use of the MPEG-21 DIDL and the OAI-PMH in an Apache Web server for which the mod_oai module was installed.

OAI-PMH GetRecord response from an Apache Web server containing a DIDL representation of a resource (a document made accessible by the Web server): modoai_GetRecord.xml
Movie 4: An interaction with a mod_oai powered Apache Web server: • Camtasia Pack & Show executable for WinTel computer; no audio; size = 7.5 Mb: modoai.exe • QuickTime movie; no audio; size = 26.3 Mb: modoai.mov

Example 8: Illustrations of the use of complex object formats with the OAI-PMH interface to an Apache Web server.

Both the discovery and preservation use cases may be addressed with mod_oai. For discovery, mod_oai offers incremental harvesting semantics with datestamp and sets (i.e. MIME types) as arguments. For preservation, mod_oai allows an entire website to be transformed into AIPs and stored for later reconstitution. The http_header metadata, either by itself or included in the oai_didl metadata format, provides complete http header information about the resource as well—information that is otherwise not available in the standard OAI-PMH usage scenario. Given the anomalies that have been reported for "Last-Modified" and "Etag" http headers in large Web crawling experiments [Clausen, 2004], it is possible that mod_oai responses can provide more accurate content gathering than standard Web crawling techniques.

Discussion

Introducing complex objects formats as metadata formats in the OAI-PMH framework yields a

robust and general solution to the resource harvesting problem. All existing OAI-PMH concepts, such as sets and "about" containers, remain available. The notion of the OAI-PMH datestamp applied to complex objects yields a reliable technique to harvest new and updated resources. The solution can be fully specified within the boundaries of the OAI-PMH, and can thus be implemented using existing, widely deployed OAI-PMH tools. Implementation boils down to specifying and implementing support for an additional metadata format.

The solution introduces an attractive archive export/ingest paradigm because the concept of transferring content between archives is approached in an application-independent and protocol-based manner, and at a more abstract level than is usually the case with mirroring solutions. Indeed, in most cases, mirroring approaches focus on complete applications or low level semantics such as files, file systems, or disk systems. The proposed approach is different in that it cleanly maps to an OAIS perspective of content transfer, as illustrated in Figure 2. In the scenario depicted, *archive1* is to be mirrored by *archive2*. Typically, *archive1* and *archive2* operate different archival environments based on different technical architectures. From a preservation perspective, such technological diversity may be a requirement rather than just a reality. From a scalability perspective, it is unrealistic to require *archive2* to implement the technological environment of *archive1* in order to be able to store safety copies of its content. Rather, it seems logical that *archive2* would prefer dealing with *archive1*'s content in the same way it deals with its own content and that from archives other than *archive1*.

In the proposed paradigm, *archive1* exposes a Dissemination Information Package (DIP) through its OAI-PMH interface. This DIP results from mapping an internal AIP (AIP 1 in Figure 2) to a complex object representation of that AIP. This complex object representation is application-neutral in the sense that it does not reflect the characteristics of the technical and architectural environment at either *archive1* or *archive2*. When transferred through the OAI-PMH, this DIP of *archive1* becomes a Submission Information Package (SIP) to *archive2*. Once transferred, *archive2* can process this SIP, and ingest it to become an AIP compliant with all other AIPs stored in its environment (AIP 2 in Figure 2). Hence the manner in which each archive internally represents resources as AIPs is of no importance. All that is important is the application-neutral complex object representation that is transferred between archives.

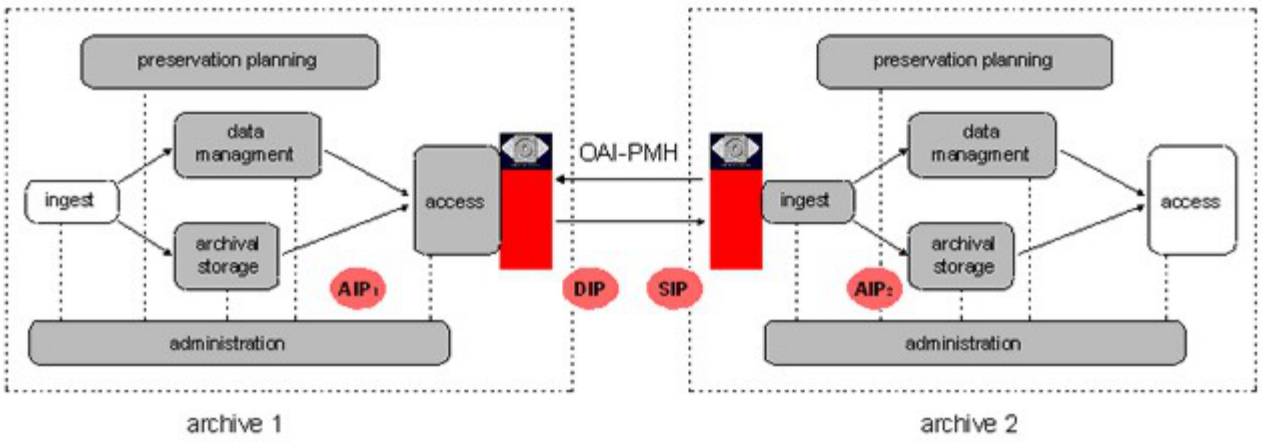

Figure 2: An OAIS perspective on content transfer between archives using the OAI-PMH

As several complex object formats exist, and many approaches exist to represent resources according to each of those formats, achieving a truly interoperable approach to transfer content using an OAI-PMH-based solution will require a specification that limits the degrees of freedom available.

When transferring content represented as complex objects, another issue that requires attention is the potential for very large records. These can occur especially if datastreams are delivered By-Value rather than By-Reference. Large records can cause problems in the implementation of the exposing

repository, as it typically may need to build the record or parts thereof in memory before transferring it. Large records can cause problems for the harvester also when it reads the records into memory and parses them. Most of the aforementioned projects have explored a technique whereby the OAI-PMH repository puts a threshold on the size of datastreams it delivers By-Value, and resorts to By-Reference delivery for files of a size that exceed that threshold. Doing so, the repository can manage its internal size-related problems. However, such an approach does not necessarily help the harvester, because its threshold may be more stringent than that of the repository it harvests from. Hence this problem domain requires further exploration. Potential solutions may include delivering bitstreams only By-Reference, or introducing some content-negotiation capability not provided by the OAI-PMH.

Another issue that requires attention is the expression of rights that apply to the resources. Indeed, concerns regarding rights can be expected to be more significant when resources are transferred instead of typical descriptive metadata. Currently, an effort is ongoing aimed at conveying rights in the OAI-PMH framework. A first result of this effort is an Implementation Guideline that specifies how to convey rights pertaining to metadata [Lagoze et al., 2004]. It is interesting to speculate about the applicability of this specification to metadata that is a complex object representation of the resource. Are rights pertaining to such metadata synonymous to rights pertaining to the resource? If so, the existing specification could be used to tackle the problem of conveying rights pertaining to the resource. If not, a planned, separate, specification aimed at expressing rights pertaining to the resource would have to be used.

Conclusions

There are a number of issues with existing approaches to resource harvesting based on the OAI-PMH. This paper presents an alternative solution, within the boundaries of the well-specified OAI-PMH. It builds on the introduction of more expressive metadata formats, complex object formats, to describe digital resources. Complex object formats allow the unambiguous distinction between an identifier of the resource and the location of a resource, and as such alleviate the lack of expressiveness that Dublin Core provides with that respect. Also, the correct interpretation of the notion of the OAI-PMH datestamp to complex object representations yields a datestamp that changes whenever a constituent of the represented resource changes. The result is a reliable trigger for incremental harvesting of resources.

This paper has also identified issues that need to be addressed in order to deploy a truly interoperable framework for resource harvesting based on the use of the OAI-PMH and complex object formats. These include reducing the degrees of freedom available in the choice and implementation of complex object formats, addressing scenarios in which large resources are to be harvested and conveying rights pertaining to harvestable resources. No doubt more issues will emerge as more parties explore the proposed approach.

We are confident that the techniques described in this paper address the need for a low-barrier and widely deployable resource harvesting solution. Our work heretofore has shown the feasibility of this solution. The fact that this solution conforms to the existing OAI-PMH specification makes its deployment simple for existing OAI-PMH implementations. As discussed, a number of issues remain for full deployment, and we hope to address these issues in the context of an effort to produce a full specification in the course of 2005. Like our previous efforts, this will involve members of the OAI community acting in a technical advisory role.

References

Bekaert, Jeroen, Patrick Hochstenbach, and Herbert Van de Sompel. 2003. "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library," *D-*

Lib Magazine, Volume 9, Number 11, November 2003. <doi:10.1045/november2003-bekaert>.

Bekaert, Jeroen, Patrick Hochstenbach, Lyudmila Balakireva and Herbert Van de Sompel. 2004. "Using MPEG-21 and NISO OpenURL for the Dynamic Dissemination of Complex Digital Objects in the Los Alamos National Laboratory Digital Library,". *D-Lib Magazine*, Volume 10, Number 2, February 2004. <doi:10.1045/february2004-bekaert>.

Clausen, Lars. 2004. "Concerning Etags and Datestamps," Fourth International Web Archiving Workshop, ECDL 2004, Bath UK. <http://www.netarchive.dk/website/publications/Etags-2004.pdf>.

Freed, N. and N. Borenstein. 1996. "RFC 2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies," November 1996. <http://www.ietf.org/rfc/rfc2045.txt?number=2045>.

Jerez, Henry, Xiaoming Liu, Patrick Hochstenbach, and Herbert Van de Sompel. 2004. "The multi-faceted use of the OAI-PMH in the LANL Repository," *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, June 7-11 2004, Tuscon, AZ, USA*. pp 11-20. <doi:10.1145/996350.996355>.

Kahn, Robert and Robert Wilensky. 1995. "A Framework for Distributed Digital Object Services. Corporation for National Research Initiatives," <http://www.cnri.reston.va.us/k-w.html>.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. "The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0". June 2002. <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2004. "OAI-PMH Implementation Guidelines: Conveying rights expressions about metadata in the OAI-PMH framework". <http://www.openarchives.org/OAI/2.0/guidelines-rights.htm>.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. "OAI-PMH Implementation Guidelines: Specification and XML Schema for the OAI Identifier Format". <http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm>.

Lossau, Norbert. 2004. "Search Engine Technology and Digital Libraries: Libraries Need to Discover the Academic Internet," *D-Lib Magazine*, Volume 10, Number 6, June 2004. <doi:10.1045/june2004-lossau>.

Maly, Kurt, Michael Nelson, and Mohammad Zubair. 1999. "Smart objects, dumb archives: a user-centric, layered digital library framework." *D-Lib Magazine*, Volume 5, Issue 3, March 1999. <doi:10.1045/march99-maly>.

Nelson, Michael. 1999. "A digital library for the National Advisory Committee for Aeronautics,"NASA/TM-1999-209127. <http://techreports.larc.nasa.gov/ltrs/PDF/1999/tm/NASA-99-tm209127.pdf>.

Nelson, Michael, Brad Argue, Miles Efron, Sheila Denn, and Maria Christina Pattuelli. 2001. "A Survey of Complex Object Technologies for Digital Libraries," NASA/TM-2001-211426. <http://techreports.larc.nasa.gov/ltrs/PDF/2001/tm/NASA-2001-tm211426.pdf>.

Osborne, Shaun. 2004. "Museums and Images JISC-FAIR Cluster Group - Images and Harvesting Issues Paper". <http://www.fitzmuseum.cam.ac.uk/htf/docs/M&I_IP_Images_jul04.doc>.

Summann, Friedrich and Norbert Lossau. 2004. "Search Engine Technology and Digital Libraries:

Moving from Theory to Practice," *D-Lib Magazine*, Volume 10, Number 9, September 2004. <doi:10.1045/september2004-lossau>.

Sidwell, C. A., P.A.D. Needham, and J.D. Harrington. 2000. "Lightening grey literature: Making the invisible visible," *New Review of Information Networking*, Volume 6, pp 121-136.

Tourte, Greg, and Andy Powell. 2004. "Encoding full-text links in the eprint jump-off page. Draft Version 1.0," <http://www.rdn.ac.uk/projects/eprints-uk/docs/encoding-fulltext-links/>.

Van de Sompel, Herbert and Carl Lagoze. 2002. "Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative," *Lecture Notes In Computer Science. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*. pp 144-157.

Van de Sompel, Herbert, Jeff Young and Thom Hickey. 2003. "Using the OAI-PMH ... Differently," *D-Lib Magazine*, Volume 9, Number 7/82, July/August 2003. <doi:10.1045/july2003-young>.

Young, Jeff. Extensible Repository Resource Locators (ERRoLs) for OAI Identifiers. <http://www.oclc.org/research/projects/oairesolver/default.htm>.

Notes

[1] DC, <http://dublincore.org/documents/dces/>

[2] MARCXML, <http://www.loc.gov/standards/marcxml/>

[3] MPEG-21, Information Technology, Multimedia Framework, "Part 2: Digital Item Declaration," ISO/IEC 21000-2:2003, March 2003.

[4] METS, <http://www.loc.gov/standards/mets/>

[5] Advanced Distributed Learning, "The Sharable Content Object Reference Model (SCORM) - Version 1.3 - WD," March 2003

[6] DARE, <http://www.surf.nl/en/themas/index2.php?oid=7>

[7] JISC FAIR, <http://www.jisc.ac.uk/index.cfm?name=programme_fair>

[8] DINI, <http://www.dini.de/>

[9] National Digital Information Infrastructure and Preservation Program, <http://www.digitalpreservation.gov/>

[10] OAIster, <http://oaister.umdl.umich.edu/o/oaister/>

[11] OCLC, <http://www.oclc.org>

[12] arXiv, <http://arXiv.org>

[13] NSDL, <http://www.nsdl.org>

[14] DC, Resource Identifier <http://dublincore.org/documents/dcmi-terms>

[15] eprints.org, <http://www.eprints.org>

[16] OA-X, <http://www.i-tor.org/oa_x/retrieving_objects/>

[17] MPEG-21, Information Technology, Multimedia Framework , "Part 3: Digital Item Identification," ISO/IEC 21000-3:2003, March 2003.

[18] International Organization for Standardization. "ISO 14721:2003. Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model (1st ed.)". 2003. Geneva, Switzerland.

[19] OAICat, <http://www.oclc.org/research/software/oai/cat.htm>

[20] OAIHarvester, <http://www.oclc.org/research/software/oai/harvester.htm>

[21] OAI Viewer, <http://www.oclc.org/research/software/oai/errol.htm>

[22] DSpace DIDL plug-in, <http://sourceforge.net/projects/didl-plug-in/>

[23] DSpace, <http://www.dspace.org>

[24] Fedora, <http://www.fedora.info/>

[25] mod_oai project, <http://www.modoai.org>

Acknowledgments

The authors would like to acknowledge:

- For their research and development on harvesting complex objects using the OAI-PMH: Lyudmila Balakireva, Jeroen, Bekaert, Mariella Di Giacomo, Henry Jerez, Xiaoming Liu, and Thorsten Schwander of the Digital Library Research and Prototyping Team of the Research Library of the Los Alamos National Laboratory
- For their efforts in the project aimed at mirroring the collection of the American Physical Society at the Los Alamos National Laboratory using the OAI-PMH: Mark Doyle and Gerard Young of the APS
- For their efforts supporting the creation of experimental plug-ins able to export complex objects from institutional repositories: Robert Tansley (DSpace & HP), Sandy Payette and Chris Wilper (Fedora & Cornell University)
- For their support of the LANL repository work and the APS mirroring project: The Library of Congress's National Digital Information Infrastructure and Preservation Program
- For their support of the mod_oai project: the Andrew W. Mellon Foundation

doi:10.1045/december2004-vandesompel